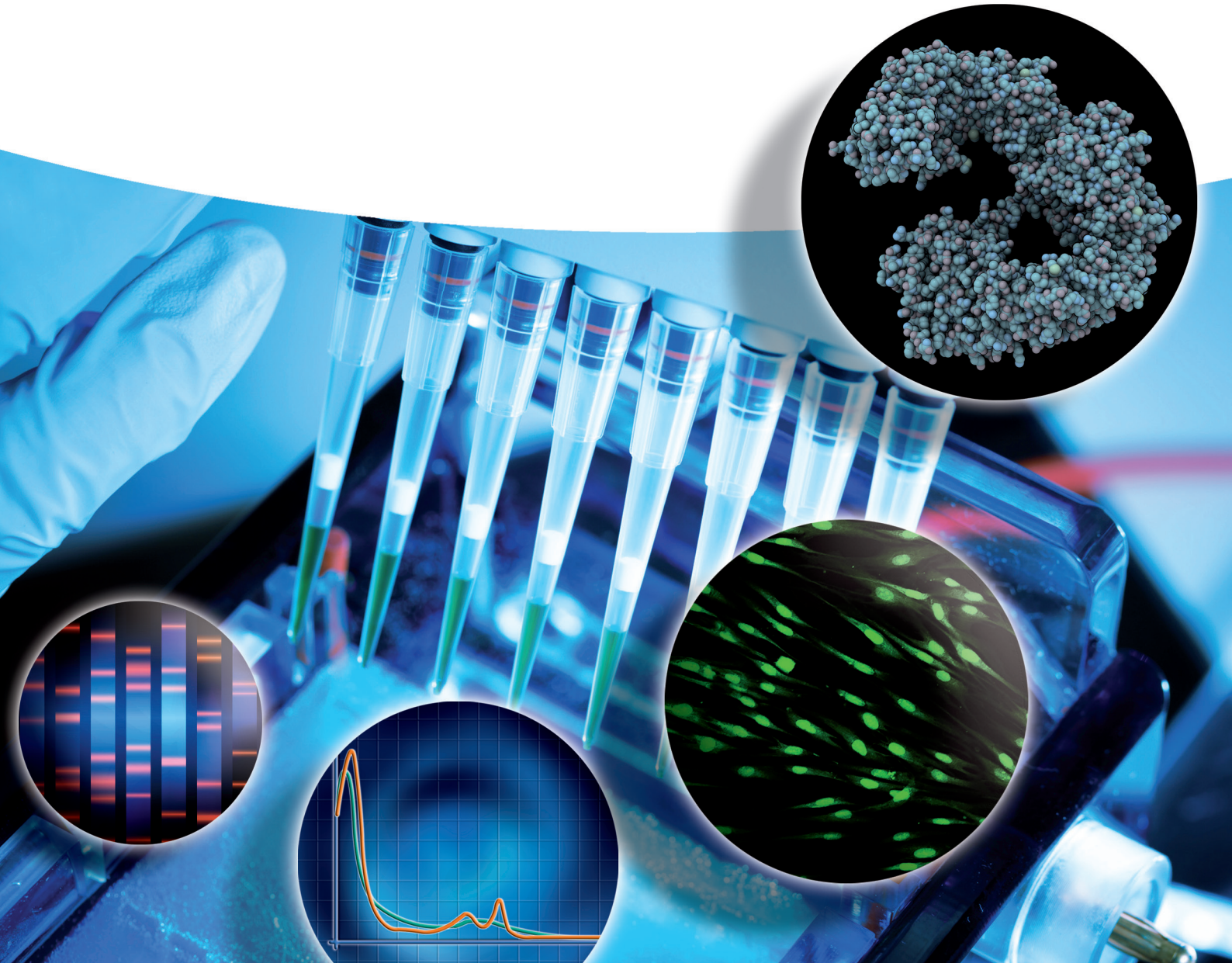
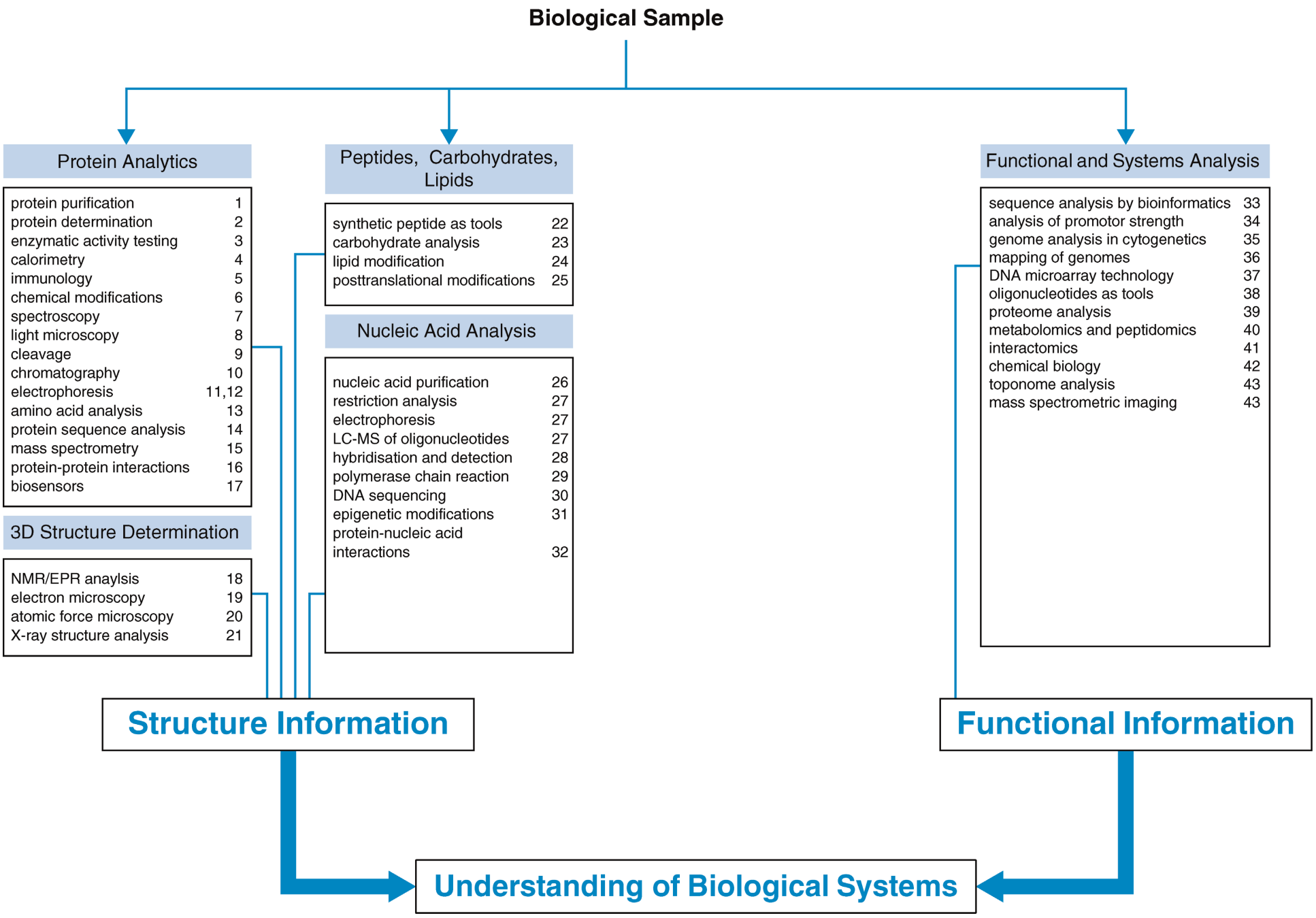
WILEY-VCH

Edited by
Friedrich Lottspeich and Joachim W. Engels

Bioanalytics

Analytical Methods and Concepts in Biochemistry and Molecular Biology




Biological Sample
Protein Analytics
protein purification 1
protein determination 2
enzymatic activity testing 3
calorimetry 4
immunology 5
chemical modifications 6
spectroscopy 7
light microscopy 8
cleavage 9
chromatography 10
electrophoresis 11,12
amino acid analysis 13
protein sequence analysis 14
mass spectrometry 15
protein-protein interactions 16
biosensors 17
3D Structure Determination
NMR/EPR anaylsis 18
electron microscopy 19
atomic force microscopy 20
X-ray structure analysis 21
Peptides, Carbohydrates, Lipids
synthetic peptide as tools 22
carbohydrate analysis 23
lipid modification 24
posttranslational modifications 25
Nucleic Acid Analysis
nucleic acid purification 26
restriction analysis 27
electrophoresis 27
LC-MS of oligonucleotides 27
hybridisation and detection 28
polymerase chain reaction 29
DNA sequencing 30
epigenetic modifications 31
protein-nucleic acid interactions 32
Functional and Systems Analysis
sequence analysis by bioinformatics 33
analysis of promotor strength 34
genome analysis in cytogenetics 35
mapping of genomes 36
DNA microarray technology 37
oligonucleotides as tools 38
proteome analysis 39
metabolomics and peptidomics 40
interactomics 41
chemical biology 42
toponome analysis 43
mass spectrometric imaging 43
Structure Information
Functional Information
Understanding of Biological Systems

Bioanalytics

Edited by

Friedrich Lottspeich and Joachim Engels

Bioanalytics

Analytical Methods and Concepts in Biochemistry and Molecular Biology

WILEY-VCH
Verlag GmbH & Co. KGaA

Editors
Dr. phil Dr. med. habil. Friedrich Lottspeich
retired from MPI for Biochemistry

Peter-Dörfler-Straße 4a
82131 Stockdorf
Germany

Prof. Dr. Joachim Engels
Goethe Universität
OCCB FB14
Max-von-Laue-Straße 7
60438 Frankfurt
Germany

Cover credit:
Background picture - ©fotolia_science photo; Circles from left to right: 1st circle - ©fotolia_T-flex; 2nd circle - Adam Design; 3rd circle - ©fotolia_fibroblasts; 4th circle - The Picture was kindly provided by Dr. Ficner.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at <http://dnb.d-nb.de>.

Print ISBN: 978-3-527-33919-8
ePDF ISBN: 978-3-527-69444-0
ePub ISBN: 978-3-527-69446-4
Mobi ISBN: 978-3-527-69447-1

Cover Design Adam Design
Typesetting Thomson Digital
Printing and Binding

Printed on acid-free paper

Table of Contents

Preface

This is a book about methods. You may ask: Why do we need a dedicated book about methods and why should I buy it? We can offer at least two good answers.

The first answer is of a theoretical nature: the method determines the quality of the scientific finding gained in that manner. Only by understanding a method, its strengths and, more importantly, its weaknesses, is it possible to estimate the general applicability of an observation or hypothesis. The development or improvement of a method is therefore a means to expand and improve the "tentative truth" generated by experimental science. Great value has been placed on describing the material critically and illuminatingly to enable the reader to engage with the material and gain a thorough understanding. This is, in our opinion, the most important reason why methods must be offered for classroom study. However, a deep and broad knowledge of methods is just as important for ongoing experimental work as it is for understanding past experiments.

The second answer is the intent – hopefully successful – of this book to make getting to know and understand these methods clear and straightforward in order to make this book an irreplaceable tool for both students and teachers. Our intent results from our conviction, backed by our experience, that today every individual, whether student, teacher, or scientist, is hopelessly overwhelmed by the large number of different techniques currently in use in biological sciences. At the same time, using these techniques is imperative.

We proudly undertook this intellectual enterprise to describe these techniques as completely as possible in an up-to-date manner. To the best of our knowledge, no English language textbook exists that is dedicated to these same goals and with the same level of coverage.

One might wonder why the most apparent reason to publish this book has not been mentioned: namely, using this book to learn, or hope to learn, methods that are needed directly for ongoing experimental work. We wish to make two things clear: This is not a "cook book". This means that after digesting a chapter the reader will not be able to go to his or her laboratory bench and apply what has just been read like a recipe – for that to be possible it will be first necessary for the reader to work through the literature relevant to the topic covered. The reader should be in a position – at least this is our goal and wish – to optimize his approach through the overview and insights acquired. As for the second point of clarification: This book does not see itself as competition for existing laboratory manuals for diverse techniques, such as protein determination or PCR. The intent is much more to use carefully coordinated and complete descriptions of the methods, using frequent cross-referencing of other chapters, either in the text or in a flanking box, to illustrate the connections between apparently unrelated techniques and to show their mutual dependencies. We believe that the reader will profit from these lessons by gaining a sense of orientation and will understand the relationship between different techniques better, or possibly appreciate them for the first time. We do not wish to conceal the fact that for us, the editors, certain methodical relationships only became clear in the course of working through some of the manuscripts. As such, this book intends to provide coverage at a higher level, more than any single method manual or a simple collection of methods could.

What is the actual content of this book? The book is titled *Bioanalytics*, which indicates that it is about analytical methods in biological sciences. This must be qualified. What are the biological sciences? Is it biochemistry or also molecular genetics, or cell and developmental biology, or even medicine? In any case, molecular biology would be included. This matter gets more complicated when one considers that modern medicine or cell biology are unimaginable without molecular biology.

This book cannot satisfy all the needs of these sciences. In addition, not all analytical methods are contained within it, instead only those that involve biological macromolecules and their modifications. Macromolecules are most often proteins, but also include carbohydrates, lipids, and nucleic acids like DNA and RNA. Special methods for the analysis of small molecular metabolites are also not included. On occasion, we have crossed over the boundaries we have set for ourselves. For example, methods for the preparation of DNA and RNA are presented, simply because they are so closely and necessarily associated with the subsequent analytical techniques. In addition, many techniques, such as electrophoresis or chromatography, can be used at

both analytical and preparative scales. For other techniques it is not easy to distinguish between preparation and analysis if one does not wish to follow the traditional division between the two based solely on the amount of material involved. Is the identification of interaction partners using the two-hybrid system an analytic method, when the final step is based on the labo intensive construction of the corresponding clones, that is, to say based on a method that, at first, does not have anything to do with investigating the interaction? Similar is the case of site-specific mutation of genes for the investigation of gene function, which first requires the construction (and not the analysis) of the mutated sequences *in vitro*. On the other hand, we intentionally omitted the description of a few techniques that are clearly preparative. The synthesis of oligonucleotides – a clearly preparative technique – and the cloning of DNA were omitted. The latter is, despite being a requirement or goal of a large number of analytical methods, not an analytical method itself. In this case our decision was easy since there are already numerous good introductions and manuals about cloning DNA. In summary, the book describes the analytical methods of protein and nucleic acid (bio)chemistry, molecular biology, and, to a certain degree, modern cytogenetics. In this context, "molecular biology" means those parts of molecular genetics and biochemistry that involve the structure and function of nucleic acids. Methods of (classical) genetics, as well as traditional cell biology, are therefore rarely, if ever, included.

We wish to emphasize that chapters that directly relate to the function of proteins and nucleic acids have been collected into a special section of the book, the "Systematic Analysis of Function." We have gone along with the shift in paradigm from traditional bioanalytics to holistic analysis approaches. In this section many topics are addressed – even though they are sometimes not entirely mature – which are on the cutting edge of science. We are aware of the fact that this area is subject to rapid change and a few aspects could in the near future, perhaps, appear to be too optimistic or pessimistic. However, we believe that discussion of the most modern techniques and strategies at this point in time covers fascinating aspects and hopefully proves to be inspiring. The increasing availability of DNA and protein sequences of many organisms is, on the one hand, the critical fundament for this systematic function analysis and, on the other hand, makes high-throughput analysis and analysis of the data increasingly important. Information gained from the genome, proteome, and metabolome is compared with *in silico* analysis, which factors in the localization and interaction between biomolecules and unites everything into complex networks. The long-term goal of completely understanding the system can surely only be reached by the incorporation of further areas of expertise that are not yet an accepted component of bioanalytics. Bioanalysts must become, and are becoming, a kind of systems biologists, more interdisciplinary, and more successful in close cooperation together with experts in the fields of informatics, system theory, biotechnology, and cell biology.

Who is this book addressed to? What has already been said provides a hint: Primarily biologists, chemists, pharmacists, physicians, and biophysicists. For some (biologists, chemists) the book will be interesting because it describes methods of their own discipline. For the second group (e.g., pharmacists, physicians, and biophysicists) the book is relevant because they can find the background and fundamentals for much of the knowledge, which they find in their own discipline. Beyond these groups, this book is dedicated to interested readers who would like to know more about the subject matter.

The material covered presumes that the user has taken at least an introductory course in the fundamentals of biochemistry or molecular genetics/gene technology, ideally both, or is in the process of doing so. We can imagine that this book would be an ideal supplement to such a course. It can and should especially be consulted when involved in experimental activities. This book is intended to be of equal value to students, teachers, and workers in these fields of science.

The organization of the material proved to be one of the most difficult aspects of putting this book together. It is almost impossible to treat the techniques used in such complex fields in the two dimensions paper offers accurately without simultaneously compromising the didactic intentions of the book. We had a choice of two approaches: a more theoretical and intellectually stringent approach or a more practically oriented approach. The theoretical approach would have been to divide the methods exclusively according to type, for example chromatography, electrophoresis, centrifugation, and so on. Under each type of method its use would be divided according to objective and by the differing types of starting materials. This approach is more logical, but harder to comprehend and unrelated to actual practice. The more practical presentation begins with the concrete problem or question and describes the method that answers the question best. This is more intuitive, but inevitably leads to redundancies. A complete deep, "multidimensional" understanding of the material is only possible after the entire book has been absorbed. The approach in this book, for the most part, follows the second, practically oriented, approach. When possible, such as in the section "Protein analysis", the methods were grouped and presented according to the topic addressed. This includes the fundamentals of instrumental techniques, which is knowledge required for the complete understanding of other sections. We approached the problem of redundancy by cross-referencing the first instance in which a method is described. Sometimes we left redundancies in place for didactic reasons. We leave it to our readers to determine if our choices represent the optimal solution to the problem of structuring the subject matter.

An overview of the presented methods and their relationships can be found on the inside back cover. This flowchart should – particularly for readers new to the topic – illustrate how one can employ the analytical approaches, from splitting open the cells down to the molecular dimensions. In the diagram, the natural turbulences of the flow are deliberately sacrificed for the sake of clarity. Hopefully, the expert reader will forgive us!

At this point we would like to explain a convention in this book, which is not in general use: the use of the terms *in vitro* and *in vivo*. To avoid misunderstandings, we explain here that we use these terms as molecular biologists usually understand them, which means using *in vitro* for "cell free" and *in vivo* for "in living cells" (*in situ* translates literally into "in place" and is used and understood as such). In contrast, pharmacologists and physicians often use the term *in vivo* to refer to experiments in animals and lack suitable terminology to distinguish between experiments conducted in cell culture and those done in test tubes. In cases

where the meaning may be unclear, we have used the precise term, “cell free”, “in living cells”, or “in animal experiments.”

This first edition in English appears some 18 years after the initial publication of *Bioanalytik* in German. We are happy about it and finally can follow the repeated wish of the scientific community and use English as the *lingua franca* of the biological sciences. The sustained interest in this book within the German-speaking community has led to our desire to make this book available to a wider international audience. To maintain the same length as the original book, despite the addition of new chapters (calorimetry, sensors, and chemical biology), we have shortened or removed other chapters. The goal was to favor current methods and to reduce method descriptions of a more historical nature. It was sometimes hard to sacrifice some cherished memories to better accommodate the current *Zeitgeist*. We would be grateful to our readers if they would point out any inaccuracies or deficiencies in our presentation, which we may have overlooked. As might be expected, this book involved a great deal of work, but was also a great deal of fun to write! We wish to thank our authors at this point, who through their conscientious and diligent work and their cooperation have been a pleasure to work with.

Last but not least, we would like to thank our publisher Wiley-VCH and its dedicated team with Waltraud Wüst and the copyeditor John Rhodes, who, with remarkable enthusiasm and tenacity, were our consistent sources of support during the realization of this book.

Joachim W. Engels and Friedrich Lottspeich
Munich and Frankfurt, January 2018

Introduction: Bioanalytics – a Science in its Own Right

In 1975, two publications by O'Farrell and Klose aroused the interest of biochemists. In their work, they showed spectacular images with thousands of neatly separated proteins, the first 2D electropherograms. At that time, a vision emerged in the minds of some protein biochemists: It might be possible to identify complex functional relationships, and ultimately to understand processes in the cell, by analyzing these protein patterns. For this purpose, however, the separated proteins had to be characterized and analyzed – a task with which the analytics at that time was hopelessly overtaxed. Completely new methods had to be developed, existing ones drastically improved, and the synergies between protein chemistry, molecular biology, genome analysis, and data processing had to be recognized and exploited, so that today's proteome analysis is on the threshold of realizing that utopian vision of more than 40 years ago.

In 1995, an international consortium with strong support from Jim Watson (HUGO; for *Hu*man *G*enome *O*rganization) decided to sequence the human genome. Even though the scientific community was initially divided on the benefits of this endeavor, those involved showed that it is possible to accomplish such a huge undertaking via international co-operation, completing it even ahead of schedule. The competition between commercial and academic participants certainly contributed to the success. Craig Venter is remembered by many for his appearance and his bold claims regarding shotgun sequencing. The major sequencing groups came from the US and Britain, such as the Sanger Institute in Cambridge, England, the Whitehead Institute in Cambridge, Massachusetts, and the Genome Sequencing Center in St. Louis. With the publication in the journal *Nature* in October 2004 the gold standard of the human genome was completed. The biggest surprise was that the actual number of genes is much lower than expected. With only about 21,000 genes, humans are nowhere near the top of the list with regard to their number of genes; being surpassed, for example, by parsley.

I.1 Paradigm Shift in Biochemistry: From Protein Chemistry to Systems Biology

The human genome project has had a fundamental impact on the entire life sciences. We now know that it is technically possible to perform fully automated high-throughput analysis in bioanalytics, and to process the enormous amounts of data that it generates. The results of the genome projects showed that predominantly data-driven research can provide fundamental insights about biology. All of this initiated a profound change from the classical, target- and function-oriented approach to biological questions to a systems-level, holistic perspective.

I.1.1 Classical Approach

Following the classical approach of the pre-genomic era, the starting point for almost every biochemical investigation was (and still is) the observation of a biological phenomenon (e.g., the alteration of a phenotype, the appearance or disappearance of enzymatic activity, the transmission of a signal, etc.). Next, an attempt is made to correlate this biological phenomenon to one or a few molecular structures, most often proteins. Once a protein has been isolated that plays a crucial role in the observed biological context, its molecular structure, including its post-translational modifications, has to be elucidated using state-of-the-art protein chemistry, so that finally the gene corresponding to this protein can be "fished". Thus, the whole arsenal of bioanalytics has to be used for the accurate analysis of an important protein. Molecular biological techniques facilitated and accelerated the analysis and validation enormously, and provided hints on the expression behavior of any proteins that were found. Physical methods such as X-ray crystallography, NMR, and electron microscopy allowed deep insights into the molecular

structures that sometimes even led to an understanding of biological processes at the molecular level. However, it was quickly recognized that biological effects are rarely explained by the action of a single protein, but are often due to the sequential actions of different biomolecules. Therefore, it was an essential step in the elucidation of reaction pathways to find interaction partners of proteins under scrutiny. When they were found, the same laborious analysis was carried out on them. It is easy to see that this iterative process was quite time-consuming so that the elucidation of a biological pathway usually took several years.

Despite its slowness, the classical approach was incredibly successful. Virtually all our current knowledge of biological processes has been gained using this strategy. It has nevertheless some basic limitations in that it is extremely difficult to elucidate network-like structures or transient interactions and to gain a complete insight into more complex reaction processes of biological systems. Another principal limitation is that the data it yields are rarely quantitative and usually reflect a rather artificial situation. This is inherent in the strategy itself, in which the complex biological system is successively broken down into modules and subunits, moving further and further away from the biological *in vivo* situation. During the many separation and analysis steps some of the initial material is inevitably lost, which will affect different proteins in different and unpredictable ways. Thus, it becomes virtually impossible to make quantitative statements, which are extremely important for a mathematical modeling of reaction processes.

I.1.2 Holistic Strategy

Encouraged by the success of the human genome project, conceptually new ways of answering biological questions began to be conceived. Instead of analytically dissecting a biological situation and then selectively analyzing the smallest units, the idea was born to view and examine the biological system as a whole (holistic, Greek *holos*, whole). The same approach is used very successfully, for example in physics, by deliberately disturbing a defined system and observing and analyzing the reaction of the system. This so-called perturbation analysis (Latin *perturbare*, to disturb) has the enormous advantage that the response of the system can be monitored without any bias, and any observed changes should be directly or indirectly due to the perturbation. This strategy is ideal for highly complex systems. It is amenable to network-like, transient and, above all, unexpected relationships, and – being based on the whole system – also very close to the real biological situation. However, to fully exploit the benefits of this strategy, the observed changes must be quantitatively measured. Due to the multitude of components in a biological system, this can be a challenge for high-throughput analytics, data processing, and advanced computing. Nevertheless, the methodological developments in bioanalytics and bioinformatics, driven and motivated by genome analysis, have reached a level that has made this kind of holistic analysis of a biological system feasible. It is seen as an essential enabling technology for systems biology, which aims to mathematically describe complex biological processes.

I.2 Methods Enable Progress

Just as two-dimensional gel electrophoresis, DNA sequencing or the polymerase chain reaction opened up hitherto unthinkable levels of knowledge about biological relationships and at the same time spurred the development in their respective fields, methodical developments regularly are at the roots of truly significant advances in science. In the last decades, the life sciences have developed rapidly and revolutionized the understanding of biological relationships. The speed of this development is closely correlated with the development of separation and analysis methods, as shown in the table below. It is almost impossible to imagine modern biochemistry without one or more of these fundamental methodological achievements.

Milestones of bioanalytical methodology

1828	urea synthesis
1866	Mendelian laws
1873	Microscopy
1890	crystallization
1894	key-lock principle
1906	chromatography
1907	peptide synthesis
1923	ultracentrifugation
1926	crystallization of urea
1930	electrophoresis
1935	phase contrast microscopy
1937	raster electron microscopy
1941	partition chromatography
1944	EPR/ESR spectroscopy
1946	radioistotope labeling
1946	NMR spectroscopy
1950	protein sequence analysis
1953	gas chromatography
1953	DNA double helix
1959	analytical ultracentrifugation
1959	PAGE
1960	hybridization of nucleic acids
1960	X-ray structure analysis
1963	solid phase peptide synthesis
1966	isoelectric focusing
1967	automated sequence analysis
1972	restriction analysis
1974	gene cloning
1974	HPLC of proteins
1975	2D electrophoresis
1975	Southern blotting
1975	monoclonal antibodies

1976	DNA sequence analysis
1981	site specific mutagenesis
1981	capillary electrophoresis
1982	transgenic animals
1982	scanning tunneling microscopy
1983	automated oligonucleotide synthesis
1985	CAT assay
1986	polymerase chain reaction
1986	atomic force microscopy
1987	MALDI MS
1987	ESI MS
1988	combinatorial chemistry
1990	phosphoimager
1990	cryo-electron microscopy
1991	yeast two-hybrid system
1993	FISH
1993	differential display
1995	proteome analysis
1995	DNA chip
1996	yeast genome sequence
1998	RNA interference
1999	STED microscopy
2004	human genome sequence
2012	CRISPR/Cas

First, separation methods were developed and their application significantly improved. Starting from the simplest separation procedures, extraction and precipitation, the conditions were created to obtain purified and homogeneous compounds via much more effective methods such as electrophoresis and chromatography. The preparation of pure substances in turn exerted an enormous development pressure on the analytical methods. It soon turned out that biomacromolecules have much more complex structures than the hitherto known small molecules. New methods had to be developed, and old ones adapted to the new requirements.

To effect a real breakthrough, the methods had to be implemented instrumentally and the instruments had to become commercially available. Since the 1950s both methods and equipment have been developed at an enormous pace. Today, they are sometimes up to 10,000 times faster and more sensitive than when they were introduced. Thanks to state-of-the-art microprocessor controls, the space requirements of the devices are also orders of magnitude lower than those of their ancestors, and their handling has similarly become easier thanks to software-assisted user guidance. While each of these tools may be quite expensive on their own, their higher throughput has led, in effect, to a tremendous cost reduction. This highly dynamic phase of method developments persists to this day. To cite one example, mass spectrometry entered biology and biochemistry, thereby enabling completely new strategies for answering biological questions, such as proteome analysis. Another important example is the success story of bioinformatics, which is used, inter alia, in the analysis of gene or protein databases and which undoubtedly has enormous potential for deployment and development. The advancement of ever-higher resolution light microscopy (near field scanning optical microscope, NFOM and confocal microscopy, 4Pi) now allows molecules to be observed in action in the cell. The well-known passage from the Bible, "because you have seen me, you have believed", is also applicable to the scientist. All of this clearly shows that we are at the beginning of a phase of transition in which analytics not only has the task of confirming the data of others as an auxiliary science, but can formulate and answer questions on its own accord as a separate, relatively complex area of expertise. Thus, analytics is changing more and more from a purely retrospective to a diagnostic and prospective science. Typical for modern analytics is the interplay of a wide range of individual processes, in which each method is limited in itself, but whose concerted action produces synergisms that can yield answers of astounding and new quality. However, in order to make this synergy possible, a scientist needs to obtain a fundamental knowledge of the areas of application, possibilities and limits of the various techniques.

I.2.1 Protein Analysis

Proteins as carriers of biological function normally must be isolated from a relatively large amount of starting material and separated from a myriad of other proteins. A purification strategy that is optimized for good yield while at the same time preserving biological activity is of utmost importance. The purification of the protein itself is still one of the greatest challenges in bioanalytics. It is often time-consuming and demands from the experimenter a substantial knowledge of the separation methods and properties of proteins. Purification is usually accompanied by spectroscopic, immunological and enzymological assays that identify and quantify proteins among a large number of very similar substances, allowing the purification process to be followed and assessed through various steps. Thorough knowledge of classical protein determination methods and enzymatic activity tests is essential, since these methods often depend on the specific properties of the protein to be measured and can be significantly influenced by contaminating substances.

Once a protein is isolated, the next step is to obtain as much information as possible about its primary structure, the sequence of its amino acid building blocks. For this purpose, the isolated protein is analyzed directly with sequence analysis, amino acid analysis and mass spectrometry. Often, the identity of the protein can be ascertained at this stage by a database query. If the protein is unknown or needs to be analyzed more closely, for example, to determine post-translational modifications, it is broken down enzymatically or chemically into small fragments. These fragments are usually separated by chromatography and some of them are fully analyzed. The determination of the full amino acid sequence of a protein with protein-chemical methods alone is difficult, laborious and expensive and is usually restricted to the quality control of recombinant therapeutic proteins. In other cases, a few easily accessible partial sequences are usually sufficient. These partial sequences are used for the preparation

of synthetic peptides, which are used in turn to generate monospecific antibodies, or oligonucleotide probes. These probes are used to isolate the gene of interest, ultimately leading to the DNA sequence through DNA analysis that is orders of magnitude faster and simpler than protein sequence analysis. This is translated into the complete amino acid sequence of the protein. However, post-translational modifications are not detected in this detour via the DNA sequence. However, since they play a decisive role in determining the properties and functions of proteins, they must be subsequently analyzed with all the available high-resolution techniques on the purified protein. These modifications can – as in the case of glycosylations – be very complex and their structure elucidation is very demanding.

Even if one knows the primary structure of a protein, has determined its post-translational modifications and can make certain statements about its folding (secondary structure), one will rarely understand the mechanism of its biological function at the molecular level. To achieve this, a high-resolution spatial structure, obtained by X-ray structure analysis, NMR, or electron microscopy, must be known. Also, the analysis of different complexes (e.g. between enzyme and inhibitor) can yield detailed insight into molecular mechanisms of protein action. Because of the high material requirements, these investigations generally take place via the detour of the overexpression of recombinant genes.

Once the entire primary structure, the post-translational modifications and possibly even the spatial structure have been elucidated, the function of a protein often still remains in the dark. Building on an intensive analysis of molecular interaction data, functional analysis is then used to deduce the functional properties from the structures of the substances studied.

I.2.2 Molecular Biology

Throughout their development, methods of biochemistry and molecular biology have mutually fertilized and supplemented each other. While molecular biology was initially synonymous with cloning, it has long become an independent discipline with its own goals, methods and results. In all molecular biological approaches, whether in basic research or in diagnostic-therapeutic or industrial applications, the experimenter deals with nucleic acids. Naturally-occurring nucleic acids exhibit a variety of forms; they can be double- or single stranded, circular or linear, of high molecular weight or short and compact, "naked" or associated with proteins. Depending on the organism, the form of the nucleic acid and the purpose of the analysis, a suitable method for their isolation is chosen, followed by analytical methods for checking their integrity, purity, shape and length. Knowledge of these properties is a prerequisite for any subsequent use and analysis of DNA and RNA.

A first approximation to the analysis of the DNA structure is provided by restriction endonuclease cleavage. Only this tool enabled the birth of molecular biology about 50 years ago. The restriction endonuclease cleavage is also the prerequisite for cloning, i.e. the amplification and isolation of uniform individual DNA fragments. It is followed by a variety of biochemical analysis methods, most notably DNA sequencing and a variety of hybridization techniques that can identify, localize, and quantify a particular, large, heterogeneous set of different nucleic acid molecules. The roughly thirty-year-old, truly Nobel Prize-worthy polymerase chain reaction (PCR) has revolutionized the possibilities of analyzing nucleic acids with a principle that is as ingenious as it is simple. Smallest amounts of DNA and RNA can be detected, quantified and amplified without cloning. The imagination of the researcher seems almost unlimited in PCR applications. Because of its high sensitivity, however, it also contains sources of error, which necessitate special caution by the user. Its evolution into a miniaturized fast and cost-effective standardized method is a good example for the lab-on-a-chip of the future.

Of course, PCR has also found its way into the sequencing of nucleic acids, one of the classic domains of molecular biology. Nucleic acid sequencing was the basis for the highly sophisticated, international human genome project. Other model organisms were also sequenced in this context. In 2010 there were about 250 eukaryotes and 4,000 bacteria and viruses sequenced, thanks in particular to modern, massively parallel sequencing methods. Many compare the human genome project with the manned flight to the moon (although it did not require similar amounts of money – the budget averaged a mere $ 200 million a year for ten years). Like similarly ambitious goals, it has led to significant technical innovations. The methods developed within the human genome project also had a major impact on biotechnology-related industries, such as medicine, agriculture or environmental protection.

An analytical method intertwined with the goals of the human genome project is the mapping of specific chromosomal regions, which is done through genetic linkage analysis, cytogenetics, and other physical processes. Mapping is done for genes (i.e. "functional units") or DNA loci which literally only exist as sequence units. A new approach has emerged with positional cloning, which used to be called reverse genetics. This "reversed" approach to traditional genetics (first gene, then function = phenotype) has already proved its worth in some cases. The most important diseases such as diabetes, cancer, heart attack, depression and metabolic diseases are each influenced by a multitude of genetic and environmental factors. Although two unrelated humans carry about 99.9% identical gene sequences, the remaining 0.1% can be crucial for the success of a therapy. Finding these differences in the gene sequences responsible for the risks offers a great opportunity to understand complex causes and processes of the disease. It will be interesting to find out which base exchanges in which positions contribute to the fact that an individual can tolerate a drug and that this drug also shows the desired effect. It is precisely this connection between the sequence on the one hand and the effect or function on the other that is the focus of functional gene diagnostics. Array diagnostics and siRNA analysis have proven to be very potent tools. Whereas the former detects the presence of the mRNA by hybridization, the latter can establish the connection between RNA and protein. The result is a high-resolution map of the human genome. siRNAs are small double-stranded RNAs (20–27-mers) that can recognize and switch off complementary mRNA. Since there are only about 21,000 genes in humans, high-throughput siRNA analysis is possible and all genes of an organism can be analyzed. For example, all genes of the nematode *Caenorhabditis elegans* have been RNAi-inhibited, leading to the first complete functional

gene mapping. Chemical biology, as a discipline of chemistry at the interface to biology, attempts to find small organic molecules for the modulation of protein interactions. Similar to the siRNA approach, cellular functions can be analyzed, but the small organic molecules have the advantage of generating fast responses that are spatially and temporally reversible. Thus, it is possible to find small molecules for all cellular targets which illuminate the physiological correlations and ultimately help to exploit new therapeutic applications.

The analysis of the linear structure of DNA is completed by determining DNA modifications, especially base methylation. They influence the structure of DNA and its association with proteins and affect a variety of biological processes. Base methylation is of particular importance for modulating gene activity. In consequence, humans with their comparatively small number of genes have the ability to regulate transcription by methylating the base cytosine. This phenomenon, known as epigenetics, is responsible for the differential expression of the genes in different cells. Because the specific modifications of the genomic DNA are lost in cloning or PCR amplification, their detection must be done directly with genomic DNA; this requires methods with high sensitivity and resolution.

I.2.3 Bioinformatics

Even before the human genome project and other sequencing projects had produced a myriad of data, the trend from wet labs to net labs has been increasing over the past thirty years; that is, the activity of some researchers increasingly shifted from the lab bench to computer-related activities. Initially, these were limited to simple homology comparisons of nucleic acids or proteins, in order to elucidate relationships or to obtain clues about the function of unknown genes. Added to this are mathematical simulations, pattern recognition and search strategies for structural and functional elements, and algorithms for weighting and evaluating the data. Databases familiar to the molecular biologist today include not only sequences but also three-dimensional structures. It is remarkable and pleasing that one has free and sometimes interactive access to this vast amount of data and their processing over the internet. This networked information structure and its management is the basis of today's bioinformatics.

I.2.4 Functional Analysis

We have already seen how bioinformatics opens up systematic functional analysis. In this section, we will cover investigations of the interactions of proteins with each other or with nucleic acids. Researchers looked at protein-DNA interactions early in the history of molecular biology after it became clear that genetic *trans*-factors are mostly DNA-binding proteins. The binding site can be characterized very precisely with so-called footprint methods. *In vivo* footprints also allow correlating the occupancy state of a genetic *cis* element with a defined process - for example, active transcription or replication. This can provide information about the mechanism of activation and also about the protein function in the cell.

Interactions between biomacromolecules can also be detected by biochemical and immunological methods, such as affinity chromatography or cross-linking methods, affinity (far-Western) blots, immunoprecipitation, and ultracentrifugation analysis. In these approaches, an unknown partner that interacts with a given protein usually needs to be subsequently identified by protein chemical methods. In genetic engineering this is easier because the interacting partner must first be expressed by a cDNA which has already been cloned. An intelligent genetic technique developed for this purpose is the two-hybrid technique, which can also be used to study interactions between proteins and RNA. It should be kept in mind, however, that the physiological significance of the identified interactions, however plausible they may appear, must be shown separately.

Protein–DNA, protein–RNA, and protein–protein interactions initiate a number of processes in the cell, including the expression of certain – as opposed to all – genes. The activity of genes expressed only in specific cell types or under specific conditions can be measured by a variety of methods, such as differential display, which is equivalent to a 1: 1 comparison of expressed RNA species. Having found genes which undergo differential expression, their *cis* and *trans* elements – in other words, the promoter and enhancer elements and the necessary transactivator proteins which effect the this regulation – can be determined. For this purpose, functional *in vitro* and *in vivo* tests are carried out.

Even though all the aforementioned analyses provide a solid insight into the specific expression of a gene and its regulation, the actual function of the gene – its phenotype – remains unknown. This is a consequence of the era of reverse genetics, in which it has become comparatively easy to sequence DNA and determine "open reading frames". Correlating an open reading frame or transcription unit with a phenotype is more difficult. Doing so requires an expression disorder of the gene of interest. This gene disorder can be introduced externally, for example by gene modification, that is, by mutagenizing the region of interest. Until about 25 years ago, site-specific mutagenesis was only possible *in vivo* in microorganisms, by the application of genetic recombination techniques. Since then, various techniques have been optimized to the point that it is possible to introduce genes modified *in vitro* into higher cells or organisms and to replace the endogenous gene. However, a disruption of the gene or of the gene function can also be achieved by other methods. In this respect, the methods of translational regulation have proven particularly useful. Replacing earlier antisense or antigenic techniques, in which oligonucleotides complementary to certain regions are introduced into the cell and inhibit the expression of the gene, the advent of RNAi in 1998 has ushered in a new era. By appropriate choice of complementary RNA any mRNA can be switched off. It is important to note that this represents not a gene knock-out, but a knock-down. Thus, crucial genes can be down-regulated without killing the organism. Instead of down-regulating, the amount of gene product can also be increased by over-expression. This has its significance for agricultural production through transgenic plants and animals. The latter methods – gene modification, antisense and RNAi technique and overexpression – have been widely used in medicine and agriculture. The reasons are as manifold as they are obvious. Transgenic animals

or plants can increase agricultural yields. Clinical expression cloning can open up new possibilities for combating malignant cells that are not recognized by the body's own immune system without the expression of certain surface antigens. The antisense and RNAi technique can be employed to suppress the activation of undesired genes, for example oncogenes. Along these lines another oligonucleotide-based method with great potential in molecular biology is the CRISPR/Cas9 technology. In prokaryotes, *C*lustered *R*egulatory *I*nterspaced *S*hort *P*alindromic *R*epeats (CRISPR) and their associated *cas* genes serve as an adaptive immune system to protect them from infection by bacteriophages. The bacterial system was adapted for use in eukaryotic cells and can now be used for precise genome engineering.

Compared to the antisense and RNAi approaches discussed above, the CRISPR/Cas9 technology is extremely efficient and easy to use. In addition, the CRISPR/Cas9 system enables the full knockout of the target gene whereas alternative methods may only result in the partial knockdown of gene expression.

However, because an organism is an infinitely more complex system than a controlled *in vitro* system or a single cell, the desired effect is not always achieved. By way of example, it should be remembered that some of the therapeutic successes of these techniques had nothing to do with nucleic acid hybridization *in vivo* but, as was later recognized, rather with a local, nonspecific activation of the immune system due to lack of methyl groups on the CpG dinucleotides of the oligonucleotides used or other protein-mediated effects. Such incidents and other, possibly less harmless complications, in our eyes lend more weight to the already existing duty of the researcher as well as the user to pay close attention to what is happening, and what can happen in their work. To that end, a solid knowledge of the available analytical methods and the interpretation of biological correlations is one of several prerequisites. This book aspires to (also) make a contribution towards that goal.

Part I

Protein Analytics

Protein Purification

1

Friedrich Lottspeich

Peter-Dörfler-Straße 4a, 82131 Stockdorf, Germany

Investigation of the structure and function of proteins has already kept scientists busy for over 200 years. In 1777 the French chemist Pierre J. Macquer subsumed under the term *Albumins* all substances that showed the peculiar phenomenon of change from a liquid to a solid state upon warming. Chicken egg white, casein, and the blood component globulin already belonged to this class of substances. Already as early as 1787 (i.e., about the time of the French Revolution) the purification of egg white-like (coagulating) substances from plants was reported. In the early nineteenth century many proteins like albumin, fibrin, or casein were purified and analyzed. It soon became apparent that these compounds were considerably more complicated than other organic molecules known at that time. The word protein was most probably introduced by the Swedish chemist Jöns J. von Berzelius in about 1838 and was then published by the Dutch Gerardus J. Mulder. Mulder also suggested a chemical formula, which at that time was regarded as universally valid for all egg white-like materials. The homogeneity and purity of these purified proteins did not correspond of course to today's demands. However, it became clear that proteins are different and distinct molecules.

At that time purification could succeed only if one could use very simple steps, such as extraction for enrichment, acidification for precipitation, and spontaneous crystallization. Already in 1889 Hofmeister had obtained chicken albumin in crystalline form. Although Sumner in 1926 could already crystallize enzymatically active urease, the structure and the construction of proteins remained unknown up to the middle of the twentieth century. Only by the development of efficient purification methods, which allowed single proteins to be isolated from complicated mixtures, accompanied by a revolution in analysis techniques of the purified proteins, was today's understanding of protein structures possible.

This chapter describes fundamental purification methods and also touches on how they can be used systematically and strategically. It is extremely difficult to look at this subject in general terms, because the physical and chemical properties of single proteins may be very different. However, this structural diversity, which in the end determines also the function of the various proteins, is biologically very meaningful and necessary. Proteins – the real tools and building materials of a cell – have to exercise a plethora of different functions.

1.1 Properties of Proteins

Size of Proteins The size of proteins can be very different. From small polypeptides, like insulin, which consists of 51 amino acids, up to very big multifunctional proteins, for example, to the apolipoprotein B, a cholesterol-transporting protein which contains more than 4600 amino acid residues, with a molecular mass of more than 500 000 Dalton (500 kDa). Many proteins are composed of oligomers from the same or different protein chains and have molecule masses up to some millions Daltons. Quite in general it is to be expected that, the greater a protein is, the more

The molar mass (M) – often wrongly called molecular weight – is not a mass but is defined as the mass of a substance divided by the amount of substance:

$$M = m/n = N_A m_M$$

The unit is g mol^{-1}

Absolute molecule mass (m_M) is the molar mass of a molecule divided by the number of molecules in one mol (= Avogadro constant, N_A).

$m_M = M/N_A$. The unit is g.

The relative molecular mass (M_r) is defined as the mass of one molecule normalized to the mass of ^{12}C (carbon 12), which by definition is equal to 12.

$$M_r = 12 \times m_{(molecule)}/m_{(^{12}C)}$$

It is dimensionless, but it has been given the "unit" Dalton (Da) (formerly atomic mass unit).

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

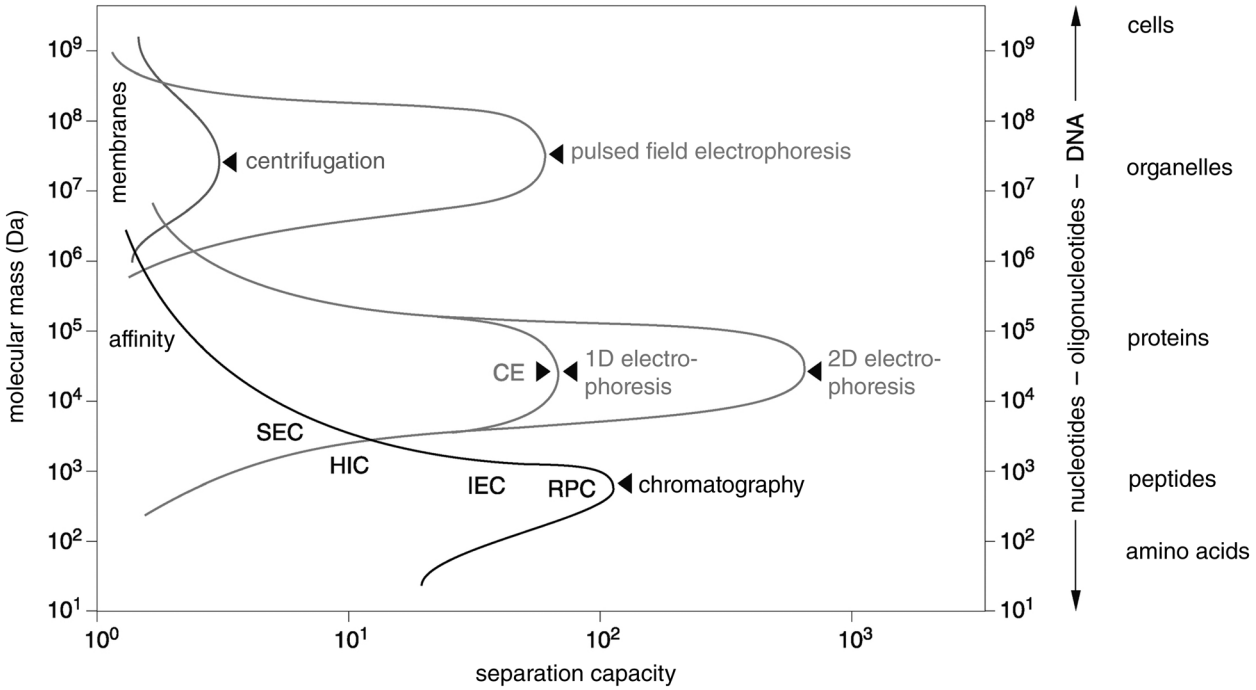


Figure 1.1 Separation methods of proteins and peptides. The separation capacity (i.e., the maximal number of compounds that can be separated in a single analysis) of the various separation methods is very different for different molecular masses of the analyte. Abbreviations: SEC, size exclusion chromatography; HIC, hydrophobic interaction chromatography; IEC, ion exchange chromatography; RPC, reversed phase chromatography; CE, capillary electrophoresis.

Dalton (Da), named after the researcher John Dalton (1766–1844), is a non-SI mass unit. One Dalton is equivalent to the atomic mass unit ($u = 1/12$ of the mass of ^{12}C) and corresponds roughly to the mass of one hydrogen atom (1.66×10^{-24} g). In biochemistry the unit kDa (1 kilodalton = 1000 Da) is very often used.

Chromatographic Separation Techniques, Chapter 10

Proteome Analysis, Chapter 39

difficultly its isolation and purification will be. This has its reason in the analytical procedures which show very low efficiencies with big molecules. Figure 1.1 shows the separation capacity (the maximum number of analytes, which can be separated under optimal condition) of individual separation techniques against the molecule mass of the analytes.

It is evident that for small molecules like amino acids or peptides some chromatographic procedures are clearly able to distinguish more than 50 analytes in a single analysis. In the area of proteins ($M_r > 104$ Da) one recognizes that of the chromatographic techniques actually only ion exchange chromatography is able to separate efficiently more complicated mixtures. In the molecular mass area of proteins electrophoretic methods are by far more efficient. That is why in proteome analysis (e.g., the analysis of all proteins of a cell), where several thousand proteins have to be separated, electrophoretic procedures (linear and two-dimensional) are very often used. From the figure is also evident that almost no efficient separation procedures exist for large molecules, for example, for protein complexes with molecular masses greater than 150 kDa, or for organelles.

The separation efficiency of a method is not always the relevant parameter in a protein purification. If selective purification steps are available the separation capacity is no longer significant and the selectivity becomes the crucial issue. Consequently, an affinity purification, which is based on the specific binding interaction of a substance to an affinity matrix, for example, an immune precipitation or an antibody affinity chromatography, has a quite low separation capacity of 1, but has an extremely high selectivity. Due to this highly selectivity a protein can easily be isolated even from a complex mixture in a one-step procedure.

With the most common purification techniques, electrophoresis and chromatography, the analytes must be present in a dissolved form. Thus, the solubility of the protein in aqueous buffer media is a further important parameter when planning a protein purification. Many intracellular proteins, located in the cytosol (e.g., enzymes), are readily soluble while structure-forming proteins like cytoskeletal proteins or membrane proteins most often are much less soluble.

Especially difficult to handle in aqueous solutions are hydrophobic integral membrane proteins, which are usually surrounded by lipid membranes. Without the presence of detergents such proteins will aggregate and precipitate during the purification.

Detergents, Section 1.8

Available Quantity The quantity available in the raw material plays a crucial role in determining the effort that must be invested for a protein purification. A protein intended for the isolation is present perhaps only as a few copies per cell (e.g., transcription factors) or as a few thousand copies (e.g., many receptors). On the other hand, abundant proteins (e.g., enzymes) can constitute percentage shares of the total protein of a cell. Overexpressed proteins of proteins are often present in in clearly higher quantities (>50% in a cell) as well as some proteins in body fluids (e.g., albumin in plasma >60%). Purification with higher quantities of a protein is usually much simpler. Especially with the isolation of rare proteins different sources of raw material should be checked for the content of the protein of interest.

Acid/Base Properties Proteins have certain acidic or basic properties because of their amino acid composition, properties that are used in separations via ion exchange chromatography and electrophoresis. The net charge of a protein is dependent on the pH of the surrounding solution. At a low pH-value it is positive, at high pH negative, and at the isoelectric point it is zero. Positive and negative charges compensate at the latter pH.

Biological Activity The purification of a protein is often complicated by the fact that a particular protein often can be detected and localized among the various other proteins only due to its biological activity and location. Hence, one must take into account at every stage of protein isolation the preservation of this biological activity. Usually the biological activity is based on a specific molecular and spatial structure. If it is destroyed, one speaks of denaturation. This often is irreversible. To avoid denaturation, one must exclude in practice the application of some procedures.

Enzyme Activity Testing, Chapter 3

The biological activity is often stable to different extents under different environmental conditions. Too high or too low buffer concentrations, temperature extremes, contacts with artificial surfaces such as glass or missing cofactors can change biological characteristics of proteins. Some of these changes are reversible: small proteins in particular are, after denaturation and loss of activity, often able to renature under certain conditions, regaining the biologically active form. For larger proteins, this is rarely the case and often results in only a poor yield. Measurement of the biological (e.g., enzymatic) activity makes it possible to monitor the purification of a protein. With increasing purity a higher specific activity is measured. In addition, the biological activity itself can be utilized for the purification of the protein. The activity often goes hand in hand with binding properties to other molecules, such as enzyme–substrate or cofactor, receptor–ligand, antibody, antigen, and so on. This binding is very specific and can be used to design affinity purifications. These are characterized by high enrichment factors and may achieve great efficiency that is difficult to obtain by other techniques.

Immune Binding, Section 5.3

Protein Purification with Tags, Section 16.2-16.4

Stability When proteins are extracted from their biological environment they are often markedly impaired in their stability. They may be degraded by proteases (proteolytic enzymes) or associate into insoluble aggregates, which almost always leads to an irreversible loss of biological activity. For these reasons, protease inhibitors are often added in the first steps of an isolation and the purification is carried out quickly and generally at low temperatures.

Protein Degradation, Chapter 9

Considering the diversity of the characteristics of proteins it immediately becomes obvious that a protein separation cannot be performed under a single schematic protocol. For a successful isolation strategy a realistic judgement of the behavior of a protein in different separation and purification methods, a minimal understanding of the solubility and charge properties of the protein to be purified, and a clear vision of why the protein is to be purified are absolutely necessary.

Goal of a Protein Purification Above all the first steps of a purification procedure, the level of purity to be aimed at and also the analytics to be used are highly dependent on the intention behind purifying a certain protein. Thus, far higher demands for cleanness must be made with the isolation of a protein for therapeutic purposes (e.g., insulin, growth hormones, or blood coagulation inhibitors) than for a protein that is used in the laboratory for structural investigations. In many cases one wants to isolate a protein only to make an unequivocal identification or to clarify some amino acid sequence segments. For such purposes usually a tiny amount of protein

(usually in the microgram range) is sufficient. With the sequence information one is able to identify the protein in protein data banks or it provides the information needed to produce oligonucleotides to isolate the gene corresponding to the protein. The protein can then be expressed in a host organism in much larger quantities (up to gram quantities) than was present in the original source (heterologous expression). Then, many of the other investigations are carried out not with the material from the natural source but with the recombinant protein. New strategic approaches to the analysis of biological questions, such as proteomics and other subtractive approaches, require completely new types of sample preparation and protein isolation. Here it is essential not to change the quantitative relations of the single proteins. A major advantage of these new strategies is that the preservation of the biological activity is no longer so important. Although each protein purification is to be regarded as a unique case, one can still can find, especially for the first purification steps, some general rules and procedures that have already been applied frequently in successful isolations; they will be discussed in detail below.

1.2 Protein Localization and Purification Strategy

The first step in any protein purification aims to bring the protein of interest into solution and remove all particulate and insoluble material. Figure 1.2 shows a scheme for different proteins. For the purification of a soluble extracellular protein, cells and other insoluble components must be removed to obtain a homogeneous solution, which can then be subjected to purification methods discussed in the following sections (precipitation, centrifugation, chromatography, electrophoresis, etc.). Sources of extracellular proteins are, for example, culture supernatants of microorganisms, plant and animal cell culture media, or body fluids such as milk, blood, urine, and cerebrospinal fluid. Often, extracellular proteins are present only in relatively low concentrations and demand as the next step an efficient concentration step.

To isolate an intracellular protein, the cells must be destroyed in a manner that releases the soluble contents of the cell and keeps the protein of interest intact. Cell disruption methods differ mainly according to cell type and amount of cells.

Membrane Proteins and other Insoluble Proteins Membrane-associated proteins are usually purified after isolation of the relevant membrane fraction. For this purpose, peripheral membrane proteins that are bound loosely to membranes are separated by relatively mild conditions, such as high pH, EDTA addition, or lower concentrations of a non-ionic detergent. This fraction of peripheral membrane proteins often can then be treated like soluble proteins. Integral membrane proteins that aggregate outside their membrane via hydrophobic amino acid sequence regions and become insoluble can only be isolated from the membrane by using high detergent

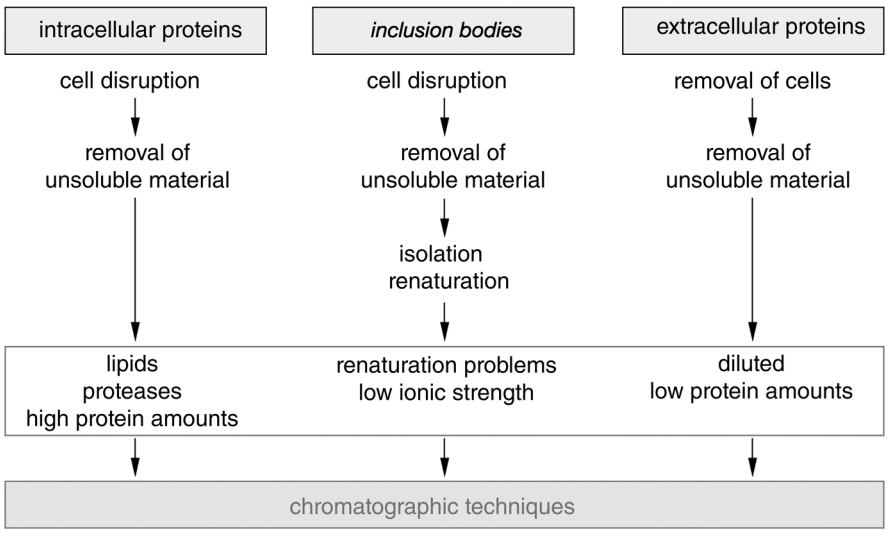


Figure 1.2 Purification scheme for different proteins. According to localization and solubility different purification steps are necessary before any subsequent selective and highly efficient purification steps.

concentrations. At present, they present probably the greatest challenge to the isolation and purification techniques.

Proteins that are insoluble in normal aqueous buffers are in general structural proteins (e.g., elastin). Additionally, they are sometimes also crosslinked via post-translationally attached modifications (e.g., functional groups). Here a first and highly efficient purification step is to remove all soluble proteins. Further steps are usually possible only under conditions that destroy the native structure of the proteins. The further processing is often carried out by cleavage of the crosslinking of the denatured proteins and the use of chaotropic reagents (e.g., urea) or detergents.

Recombinant Proteins A special situation occurs in the production of recombinant proteins. A rather simple purification is possible by the expression of recombinant proteins in inclusion bodies. These are dense aggregates of the recombinant product, which are present in a non-native state and are insoluble, because the protein concentration is too high, or because the expressed protein in the host environment cannot be correctly folded, or because the formation of the (correct) disulfide bonds in the reducing environment inside the host is not possible. After a simple purification by differential centrifugation (Section 1.5.2), in which the other insoluble cell components are removed, the recombinant protein is obtained in a rather pure form. However, it still needs to be converted into the biologically active state by renaturation.

Protein Interaction, Section 16.2–16.4

When the expression of recombinant proteins does not result in inclusion bodies, the protein is present in a soluble state inside or outside of the cell, depending on the vector. Here, further purification is similar to the purification of naturally occurring proteins but with the advantage that the protein to be isolated is already present in relatively large amounts.

Recombinant proteins can be easily isolated by using specific marker structures (tags). Typical examples are fusion proteins in which at the DNA level the coding regions for a tag structure and the desired protein are ligated and expressed as a single protein. Such fusion proteins often can be isolated in a rather pure form in a one-step procedure on applying a specific antibody affinity chromatography against the tag structure. Examples are GST fusion proteins with antibodies against GST or biotinylated proteins using avidin columns. Another frequently used tag-structure is multiple histidine residues, which are attached to the N- or C-terminal end of the protein chain and are easy to isolate by immobilized metal affinity chromatography (IMAC).

Immobilized Metal Affinity Chromatography, Section 10.4.8

1.3 Homogenization and Cell Disruption

To purify biological components of intact tissues, the complex cell associations must be disrupted in a first step by homogenization. The result is a mixture of intact and disrupted cells, cell organelles, membrane fragments, and small chemical compounds derived from the cytoplasm and from damaged subcellular compartments. Since the cellular components are transferred to a non-physiological environment, the homogenization media should meet several basic requirements:

- protection of the cells from osmotic bursting,
- protection from proteases,
- protection of the biological activity (function),
- prevention of aggregation,
- minimal destruction of organelles,
- no interference with biological analyses and functional tests.

Normally this is done by isotonic buffers at neutral pH. Often, a cocktail of protease inhibitors is added (Table 1.1).

If you want to isolate intracellular organelles, such as mitochondria, nuclei, microsomes, and so on, or intracellular proteins, the still intact cells have to be disrupted. This is accomplished by mechanical destruction of the cell wall. This procedure releases heat of friction and therefore has to be carried out with cooling. The technical realization of the disruption process varies depending on the starting material and location of the target protein (Table 1.2).

Table 1.1 Protease inhibitors.

Substance	Concentration	Inhibitor of
Phenylmethylsulfonyl fluoride (PMSF)	0.1–1 mM	Serine proteases
Aprotinin	0.01–0.3 µM	Serine proteases
ε-Amino-*n*-caproic acid	2–5 mM	Serine proteases
Antipain	70 µM	Cysteine proteases
Leupeptin	1 µM	Cysteine proteases
Pepstatin A	1 µM	Aspartate proteases
Ethylenediaminetetraacetic acid (EDTA)	0.5–1.5 mM	Metalloproteases

For very sensitive cells (e.g., leukocytes, ciliates) repeated pipetting of the cell suspension or pressing it through a sieve is sufficient to achieve a disintegration by surface shear forces. For the slightly more stable animal cells, the shear forces are generated with a glass pestle in a glass tube (Dounce homogenizer). These methods are not suitable for plant and bacterial cells.

- Cells that have no cell wall and are not associated (e.g., isolated blood cells) can be broken osmolytically by being placed in a hypotonic environment (e.g., in distilled water). The water penetrates into the cells and causes them to burst. In cells with cell walls (bacteria, yeasts) the cell walls must be treated enzymatically (e.g., with lysozyme) before an osmolytic digestion can succeed. Such exposure is very gentle and is therefore particularly suitable for the isolation of nuclei and other organelles.
- For bacteria repeated freezing and thawing is often used as a disruption method. By changing the aggregate state the cell membrane is deformed so that it breaks and the intracellular content is released.
- Microorganisms and yeasts can be dried at 20–30 °C in a thin layer for two to three days. This leads to destruction of the cell membrane. The dried cells are then ground in a mortar and can be stored at 4 °C if necessary also for longer periods. Soluble proteins can be extracted with an aqueous buffer from the dry powder in a few hours.
- With cold, water-miscible organic solvents (acetone, –15 °C, ten-times volume) cells can be quickly drained, with the lipids extracted into the organic phase, and thus the cell walls are destroyed. After centrifugation, the proteins remain in the precipitate, from where they can be recovered by extraction with aqueous solvents.
- With stable cells such as plant cells, bacteria, and yeasts a mortar and pestle can be applied for cell disruption, although larger organelles (chloroplasts) may be damaged. The addition of an abrasive (sea sand, glass beads) facilitates the disruption.
- For larger quantities, a rotating knife homogenization can be used. The tissue is cut by a rapidly rotating knife. As this produces considerable heat a way of cooling should be present. For small objects such as bacteria and yeasts, the efficiency of the pulping process is significantly improved by the addition of fine glass beads.
- Vibration cell mills are used for a relatively harsh disruption of bacteria. These are lockable steel vessels in which the cells are vigorously shaken with glass beads (diameter 0.1–0.5 mm). Again, the heat generated must be dissipated. Cell organelles can be damaged in this decomposition method.
- Rapid changes in pressure break cells and organelles in a very efficient manner. Therefore, strong pressure changes are produced in the suspension of a cell material with ultrasonic waves in the frequency range 10–40 kHz through a metal rod. Since in this method much heat is released, only relatively small volumes and short sound pulses with a maximal duration of 10 s should be applied. DNA is fragmented under these conditions.
- In a further disruption method that is particularly suitable for microorganisms, up to 50 ml of a cell suspension are pressed through a narrow opening (<1 mm), where the shear forces destroy the cells (French press).

Table 1.2 Biological starting materials and disruption methods.

Material	Disruption method	Comments
Bacteria		
Gram positive	Enzymatic with lysozyme EDTA/Tris French press	Peptidoglycan cell wall renders cell wall permeable
Gram negative	Cell mill with glass beads freeze–thaw ultrasonic	Mechanical disruption of cell wall not suitable for large amounts because of local overheating
Yeast	Autolysis French press mechanical with glass beads enzymatic with zymolase	24–28 h with toluene several times, since rather inefficient good efficiency add protein inhibitors
Plants	Waring blender + reducing agents + phenol oxidase inhibitors + protease inhibitors	Disruption of the cell wall dithiothreitol polyvinylpyrrolidone high content of proteases in plants
Fibrous tissues	Grind in liquid nitrogen	Cold homogenization buffer
Non-fibrous tissues	Grind, eventually after drying	
Higher eukaryotes		
Cells from suspension culture	Osmolysis with hypotonic buffer press through a sieve repeated pipetting of the suspension	Very damageable cells add protease-inhibitors
Fibrous tissues	Chop into small pieces Dounce Homogenizer	
Muscle tissues	Chop into small pieces with a meat grinder	Difficult to disrupt

Source: adapted from: Deutscher, M.P. (ed.) (1990) *Guide to Protein Purification*, Methods in Enzymology, vol. 182, Academic Press.

Depending on the objectives, proteins of interest in soluble form must be subjected to further purification steps. For this purpose, the homogenate is usually roughly separated into different fractions by differential centrifugation (Section 1.5.2)

1.4 Precipitation

The precipitation of proteins was one of the first techniques used for the purification of proteins (salting out of proteins was done for the first time over 130 years ago!). The method is based on the interaction of precipitating agents with the proteins in solution. These agents may be relatively non-specific resulting in the precipitation of practically all proteins from a solution. This is one way to obtain the total proteins from a cell lysate and is used frequently in the first steps of a purification process to obtain and concentrate the whole protein content of a cell. The precipitation can also be carried out in such a way that a fractionation of the components of a solution is possible. An example of this is the Cohn-fractionation of plasma, which was formulated in 1946 and still is used for plasma protein production on a large scale. Blood plasma is mixed with increasing amounts of cold ethanol and the proteins that are insoluble at the respective ethanol concentration are precipitated by centrifugation. However, with the exception of the precipitation of antigens with antibodies, precipitation is not protein specific and therefore can only be used for a rough pre-fractionation of protein mixtures.

Depending on the problem and starting material, the precipitation can be-performed under different conditions. As well as the efficiency of the precipitation in itself, further aspects also be considered:

- Is the biological activity affected by the precipitant and precipitation conditions?
- Under what conditions can the precipitant be removed?

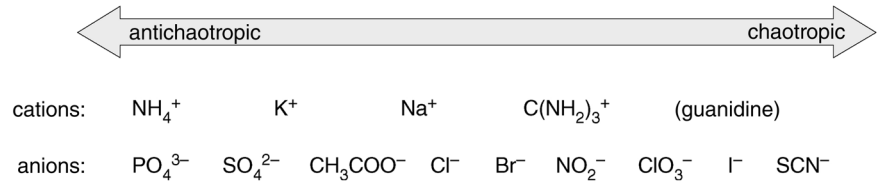


Figure 1.3 The Hofmeister series.

The ability of a salt to precipitate proteins is described in the so-called Hofmeister series (Figure 1.3). Salts further to the left (so-called-antichaotropic or cosmotropic salts) are particularly good and gentle precipitants. They increase hydrophobic effects in the solution and promote protein aggregation via hydrophobic interactions. Salts further to the right (chaotropic salts) reduce hydrophobic effects and keep proteins in solution. The oldest method used to precipitate proteins is the "salting out" of proteins by adding ammonium sulfate. The concentration of the proteins prior to precipitation should be about 0.01–2%. Ammonium sulfate is particularly suitable since in concentrations above 0.5M it also protects the biological activity of sensitive proteins. It is easily removed from the proteins (dialysis, ion exchange) and is also inexpensive. Thus it may be used already in the very first purification steps for the precipitation from larger volumes. Ammonium sulfate is added usually under controlled conditions (temperature, pH) in portions to the protein solution, whereby a fractional precipitation and thus an accumulation of the protein of interest can be achieved. Notably, a complete precipitation can take several hours. Ammonium sulfate precipitates are usually tight and easy to centrifuge (100*g*, cf. Section 1.5). The only major drawback of ammonium sulfate relates to the precipitation of proteins that require calcium for their activity/structure because calcium sulfate is virtually insoluble and thus it will be removed from the proteins. Therefore, these proteins have to be precipitated with other salts (e.g., acetates).

Precipitation using Organic Solvents For over 100 years it has been known that proteins can be precipitated with cold acetone or short-chain alcohols (mainly ethanol). Long-chain alcohols (greater than C_5) have a low solubility in water and are not suitable for precipitation. For the choice of organic precipitant or the optimal temperature no general rules can be given. Ethanol has proven particularly useful in the precipitation of plasma proteins. For protein solutions that contain lipids, acetone is often used, since in addition to the precipitation of the proteins at the same time the lipids are extracted. To avoid too high a local concentrations of the organic solvent, which may result in the denaturation of the proteins, the solvent should be added slowly. Good cooling and slow addition are also useful, since by the addition of organic solvent (e.g., ethanol to water) heat may be generated, which could lead to unwanted denaturation. The precipitate can be pelleted by centrifugation (see below) and again dissolved in aqueous buffers. In a frequently used protocol for acetone precipitation a fivefold volume excess of cold acetone (–20 °C) is added to the protein solution and incubated overnight at –20 °C. Then, it is centrifuged for 30 min at 20 000*g*. This precipitation usually provides excellent results even for very small quantities of protein. The recovery of protein (yield of the precipitation) must be monitored with analytical methods (SDS-gel electrophoresis or activity tests, etc.).

Precipitation using Trichloroacetic Acid One method commonly used to precipitate proteins from solutions is the precipitation with 10% trichloroacetic acid, wherein a final concentration of 3–4% should be reached. After centrifugation, the precipitate is resuspended in the desired buffer for further purification steps, wherein the pH of the solution should be checked. This method denatures the proteins and is therefore primarily used for the concentration for gel electrophoretic separation or prior to enzymatic cleavages. The minimum sample concentration should be 5 µg ml^{-1}.

Precipitation of Nucleic Acids Protein solutions of cell lysates, especially bacteria and yeasts, contain a large amount of nucleic acids (DNA and RNA) which can interfere with protein purification and therefore usually has to be removed. Because nucleic acids are highly negatively charged polyanions they can be precipitated with strong alkaline substances (e.g., polyamines, polyethyleneimines or anion exchange resins) or very basic proteins (protamines). Optimization of the nucleic acid precipitation – and washing conditions – must avoid also precipitating the

proteins of interest, either by the precipitating agents used or due to binding to the precipitated nucleic acids (e.g., histones, ribosomes).

1.5 Centrifugation

Centrifugation is not only one of the oldest techniques for the removal of insoluble constituents, it is also used for cell fractionation and isolation of cell organelles. It is based on the movement of particles in a liquid medium by centrifugal forces. The central part of a centrifuge, the rotor, serves to receive the sample cup and is driven by a motor to high rotational speed. There are various construction designs of rotors, such as fixed-angle rotors, vertical or swinging bucket rotors (Figure 1.4), which are available in various sizes and materials. They allow separations of a few microliters to several liters and can be operated according to the task in hand with various adjustable rotational speeds. Usually, cooled centrifuges are used for working with biological materials. High-speed centrifuges, ultracentrifuges, are operated consistently under vacuum to avoid frictional heat that occurs at high speeds due to air resistance. In the operation of

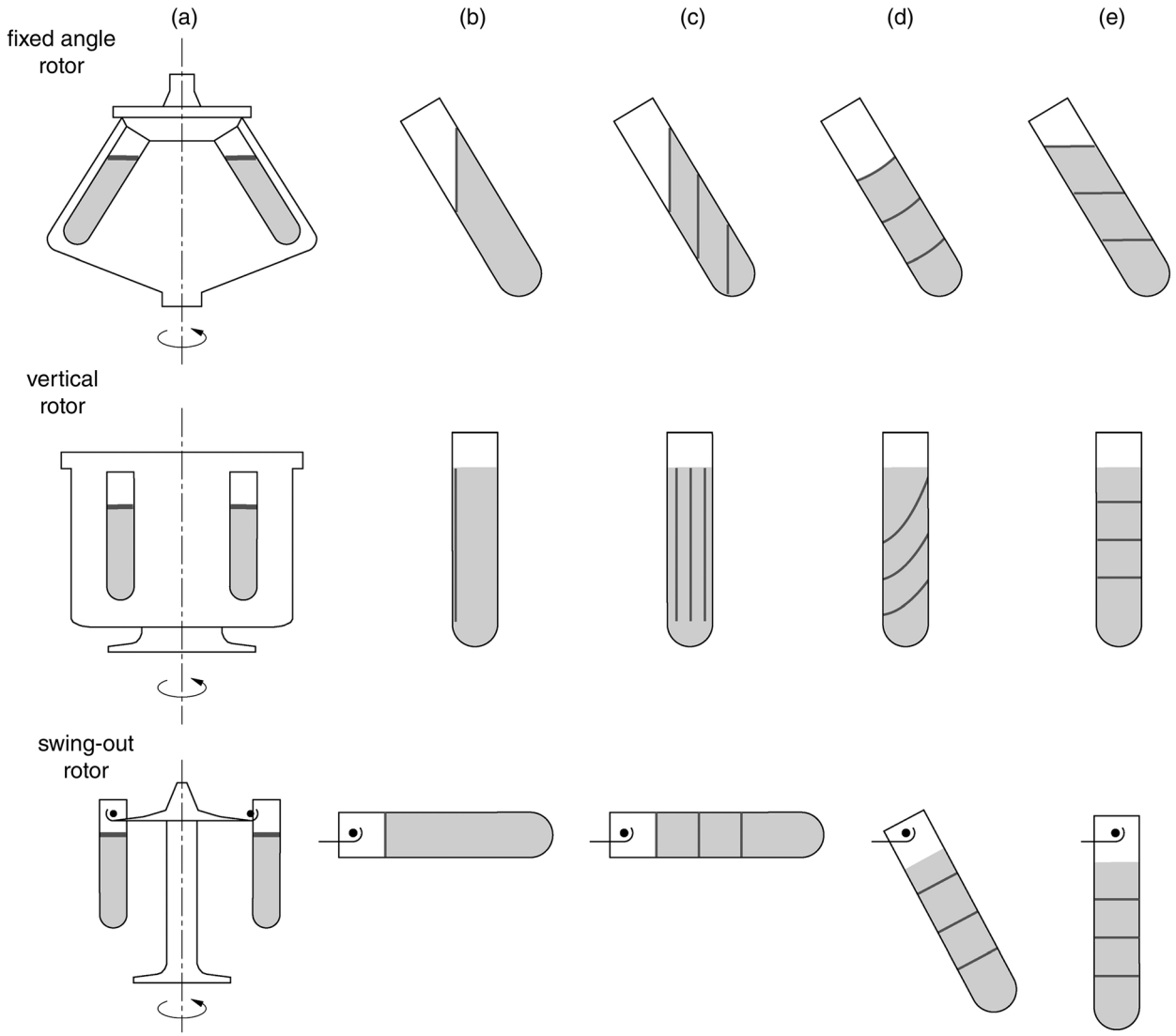


Figure 1.4 Rotors for centrifugation. Fixed angle rotor, vertical rotor, and swing-out rotor – loading conditions (a); under centrifugation at start of separation (b); during separation (c); slowing down (d) and finish of centrifugation (e); protein containing fractions in dark grey.

centrifuges certain safety precautions must be followed, in particular opposite sample containers must be well balanced to avoid any imbalance that could destroy the centrifuge.

1.5.1 Basic Principles

The physical principle of centrifugation is a separation according to size and density. On a particle being moved at a constant angular velocity around an axis of rotation, a centrifugal force is acting, which accelerates the particles outwardly. The acceleration (B) is dependent on the angular velocity and on the distance r from the rotation axis:

$$B = \omega^2 r \tag{1.1}$$

The acceleration is related to gravitational acceleration g (981 cm s^{-2}) and expressed as relative centrifugal force (RCF) in multiples of the acceleration due to gravity (g):

$$\mathrm{RCF} = \omega^2 r/981 \tag{1.2}$$

The relationship between the angular velocity and the rotational speed in rotations per minute (rpm) is given by:

$$\omega = \pi \times \mathrm{rpm}/30 \tag{1.3}$$

then, by substitution:

$$\mathrm{RCF} = 1.118 \times 10^{-5} r(\mathrm{rpm})^2 \tag{1.4}$$

Notably, during centrifugation usually the distance of the particle from the axis of rotation and thus the RCF changes. Therefore, for the calculations the mean distance is often used.

The sedimentation rate of spherical particles in a viscous fluid is described by Stokes' law, given below, where, ν is the sedimentation rate, g is the relative centrifugal acceleration, d is the diameter of the particle, and ρ_p and ρ_m are the density of the particle or the liquid, respectively, and η is the viscosity of the medium. The sedimentation rate increases with the square of the particle diameter and the difference in densities between particle and medium, and decreases with the viscosity of the liquid. Now, if the sedimentation takes place in a medium such as, for example, 0.25 M sucrose, which is less dense than all the particles, and also has a low viscosity, the diameter of the particles is the dominant factor for the sedimentation velocity.

Stokes' law:

$$v = d^2(\rho_p - \rho_m)g/18\eta \tag{1.5}$$

The sedimentation coefficient (s) is the sedimentation velocity under geometrically specified conditions of the centrifugal field. It is specified in Svedberg units (S):

$$s = v/r\omega^2 \tag{1.6}$$

1 S corresponds to 10^{-13} s. This order of magnitude is valid for different biological molecules. The Svedberg unit of a biomolecule sometimes appears in its name (e.g., 18S-rRNA), which then allows a conclusion to be made as to the size of the particle. Table 1.3 gives the size and conditions for the purification by centrifugation of cells and a few cellular compartments. A good overview is given by the representation of the particles in a density/sedimentation coefficient diagram (Figure 1.5) or in a density/g-values chart. From Stokes equation, the various techniques of centrifugation can be easily understood.

1.5.2 Centrifugation Techniques

Differential Centrifugation Differential centrifugation exploits the different sedimentation rates of different particles. It is carried out in a fixed-angle rotor and assumes that the sedimentation rates are sufficiently different. The large and heavy nuclei sediment relatively quickly (1000g for 5–10 min) and can be found even at low centrifugation speed in the precipitate (pellet). At higher RCF mitochondria (10 000g, 10 min) and microsomes (100 000g for 1 h) sediment. However, the individual fractions are not pure because slow particles that were geometrically close to the bottom of the centrifuge tube are contaminated by fast particles that were near the surface and must travel a longer distance. Differential centrifugation is not only used

Table 1.3 Typical density, particle diameter, and *S*-values of biological materials.

Particle	Diameter (μm)	Density (g cm^{-3})	*S*-Value	Sedimentation in sucrose gradient at
Cells	5–15	1.05–1.2	10^7–10^8	1000*g*/2 min
Nuclei	3–12	>1.3	10^6–10^7	600*g*/15 min
Nuclear membrane	3–12	1.18–1.22		1500*g*/15 min
Plasma membrane	3–20	1.15–1.18		1500*g*/15 min
Golgi-apparatus	1	1.12–1.16		2000*g*/20 min
Mitochondria	0.5–4	1.17–1.21	1×10^4–5×10^4	10 000*g*/25 min
Lysosomes	0.5–0.8	1.17–1.21	4×10^3–2×10^4	10 000*g*/25 min
Peroxisomes	0.5–0.8	1.19–1.4	4×10^3	10 000*g*/25 min
Microsomes				100 000*g*/1 h
Endoplasmic reticulum	0.05–0.3	1.06–1.23	1×10^3	150 000*g*/40 min
Ribosomes		1.55–1.58	70–80	
Soluble proteins	0.001–0.01	1.2–1.7	1–25	

for the enrichment of particles but also for concentration. Thus, for example, from one liter of bacterial cell culture the cells can be pelleted by centrifugation in 15 min at 2000*g* and then can be resuspended in a smaller volume.

Zonal Centrifugation If the sedimentation rates of molecules do not differ sufficiently, the viscosity and density of the medium can be used to generate selectivity. In the zonal centrifugation a preformed flat density gradient, mostly from sucrose, is used and the sample layered over the gradient (see below). The particles which at the beginning of the centrifugation – in contrast to differential centrifugation – are present in a narrow zone are now separated according to the sedimentation velocity. The density gradient, in addition to the minimization of convection, also has the effect that at increasing density and viscosity those faster particles are slowed down that would otherwise sediment with the increasing RCF caused by increasing distance from the rotor axis. This gives an approximately constant rate of sedimentation of the particles. Zonal centrifugation, which is usually carried out at relatively low speeds with swinging-or vertical

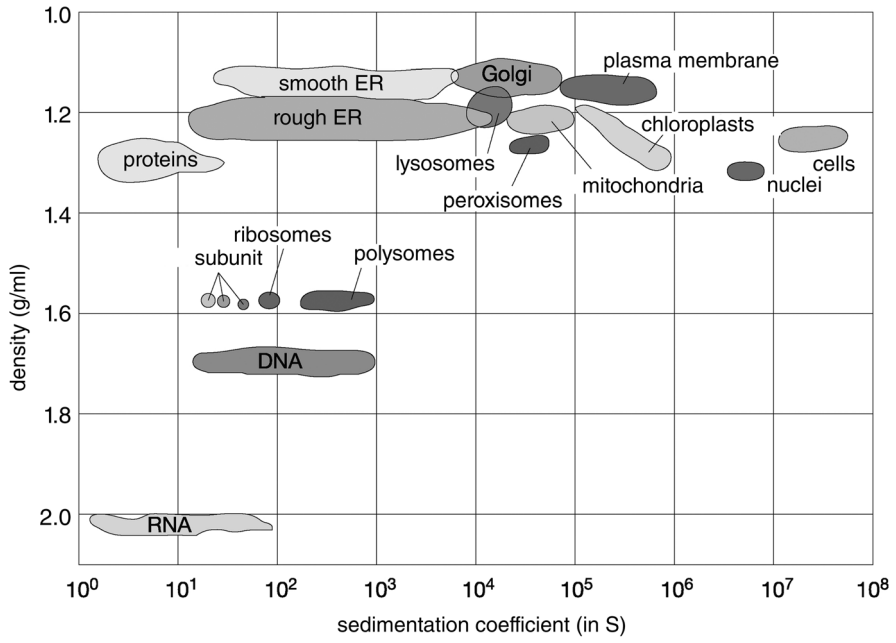


Figure 1.5 Density and sedimentation coefficients of some cell compartments. The figure shows the distribution of different cell components in terms of their density and their sedimentation coefficients. ER, endoplasmic reticulum.

rotors, is an incomplete sedimentation; the maximum density of the medium must not exceed the lowest density of the particles. The centrifugation is stopped before the particles pellet.

Isopycnic Centrifugation The previously discussed techniques of differential and zonal centrifugation are especially suitable for the separation of particles that differ in size. These techniques are not well suited for particles having a similar size but different densities. For these cases, isopycnic centrifugation (also known as sedimentation equilibrium centrifugation) is used. Here centrifugation is performed for long periods at high speed in a density gradient until equilibration. According to Stokes' equation, particles remain in the floating state when their density and the density of the surrounding medium are equal ($v = 0$). Particles in the upper part of the centrifuge tube sediment until they reach the state of suspension and cannot sediment further because the layer below has a greater density. The particles in the lower region rise accordingly up to the equilibrium position. In this type of centrifugation, the gradient density must exceed the density of all particles to be centrifuged.

Density Gradient To generate the density gradient, which can be continuous or discontinuous (in stages), various media are used, which have been found for the different application areas as appropriate:

- **CsCl solutions** can be prepared with densities up to 1.9 g ml^{-1}. They are of very low viscosity, but have the drawback of high ionic strength, which can dissociate some biological materials (chromatin, ribosomes). In addition, CsCl solutions have high osmolality, which makes them unsuitable for osmotically sensitive particles such as cells. CsCl gradients are particularly suitable for the separation of nucleic acids.
- **Sucrose** is often used for the separation of subcellular organelles in zonal centrifugation. The inexpensive and easy to prepare solutions are nonionic and relatively inert to biological materials. The low density of isotonic sucrose solutions (<9% w/v) often prevents the centrifugation of cells. In isopycnic centrifugation the high viscosity of highly concentrated sucrose solutions leads to poor resolution.

Because of the high osmolality of natural sucrose, high molecular weight polysaccharides such as glycogen, dextran, or synthetic polysaccharides such as Ficoll are used to generate the gradients. Although these polysaccharides have a better osmolality than sucrose, their high viscosity leads to longer centrifugation times and poorer separations. The polysaccharides are used for zonal centrifugation and isopycnic centrifugation. After centrifugation one may dispose of the polysaccharides by dilution and subsequent precipitation of biological particles.

Colloidal silica particles coated with polymer may also be used as density gradient media. They show lower osmolality, but higher viscosity than the cesium salts. Percoll consists of polyvinylpyrrolidone (PVP) coated silica particles about 15–30 nm in diameter, which are available as a suspension with a density of 1.130 g ml^{-1}. During centrifugation density gradients form quickly by itself due to the colloidal nature of Percoll. The profile, however, changes during centrifugation. An isotonic density gradient can be formed by addition of sucrose or salts to Percoll.

Today a class of iodinated media is used for the formation of a stable, inert, and non-toxic density gradient; they were originally developed as contrast media for X-ray structure analysis. The most widely used representative of these media is Nycodenz, which is particularly suitable for centrifugation of cells and membrane bound particles.

Fractionation of the Separated Bands After centrifugation the separated bands have to be recovered from the centrifuge vessel. If discontinuous gradients were used, the fractions of interest are sometimes visible to the density limits and can be gently aspirated with a Pasteur pipette. For a continuous gradient in which the fractions are often not clearly visible, fractionation is performed by making a hole in the bottom of the centrifuge tube and then collecting the gradient dropwise in sample containers. Another method is to introduce a dense liquid to the bottom of the tube which can displace the entire gradient upwards. With a suitable device at the top of the tube, the gradient can be delivered into a fraction collector. Then a solution that exceeds the density of the gradient is pumped through the glass tube to the bottom of the centrifuge vessel. The gradient is lifted and collected at the fraction collector.

1.6 Removal of Salts and Hydrophilic Contaminants

In the course of protein purification often solutions are obtained that are inappropriate in ionic strength or buffer composition for the next purification step. For example, after hydrophobic interaction chromatography the salt content of the sample is too high for direct subsequent ion exchange chromatography. In many cases, through good planning of the purification strategy, it will be possible to avoid additional steps for desalting. Especially in the last purification step prior to further analytical investigations the sample should be salt-free by using a volatile solvent system. On the other hand, there are several ways of buffer changing and desalting, which afford good yields at least with highly soluble proteins.

Chromatography, Chapter 10

Dilution Often it is very easy to reach the ionic strength required for the next purification step by simply diluting the sample with distilled water. For the next step in the purification usually a concentration step, such as ion exchange chromatography or affinity purification, is necessary. If dilution is insufficient to lower the salt concentration in the desired manner, desalting techniques need to be applied, which will be described next.

Dialysis The longest known desalting method is dialysis. The most important part of dialysis is the dialysis membrane, where small molecules can diffuse freely, while larger molecules are retained. There are various membranes available, which differ mainly in pore size (molecular weight cut-off threshold, called the cut-off value). The cut-off value usually indicates the molecular weight of proteins that are excluded by 90% of the membrane. These values are determined with dextrans or globular proteins and are expressed in the membrane description. But, in addition, the hydration and the charge of a protein also play an important role in passage through the membrane. The cut-off value does not give a sharp molecular weight cut-off but may only give a clue as to which molecular size can pass through the membrane relatively unhindered. The protein solution is placed in a dialysis tube, consisting of a dialysis membrane. It is recommended that prior to dialysis the tubing, which is usually made of regenerated cellulose and contains significant amounts of heavy metals, is boiled for a few minutes in distilled water and rinsed with distilled water extensively to remove impurities. Since the volume of the sample solution can significantly increase due to water migration during dialysis, the dialysis tube should only be filled to no more than two-thirds full. The so-filled dialysis tube, largely free of air and sealed by nodes at both ends, is suspended in a beaker with the desired buffer. The rate of diffusion through the membrane is determined by the concentration gradient of the diffusible particles, the diffusion constants of these particles, the membrane surface, and the temperature. For effective desalting the buffer has to be stirred and changed several times to obtain the largest possible concentration gradient across the membrane surface. For the stability of proteins, dialysis is normally performed in a cold room. The progress of the desalting can be checked by measuring the conductivity of the buffer. Normally, two-to-three buffer exchanges with a 4–6 h equilibration period each are sufficient.

In an extensive desalting (dialysis against water) it must be noted that proteins can partially precipitate because of a low ionic strength. The precipitates may be centrifuged and often can be dissolved in a small volume of a solution with a slightly higher ionic strength.

Dialysis is a very simple, relatively slow technique that is limited to very small amounts of sample (micrograms) by adsorption losses. For these small amounts and small volumes (<500 µl) instead of dialysis tubes various special constructions are used that have a small sample chamber of a few microliters (e.g., Eppendorf vessels or the lid of the Eppendorf vessel) and can be sealed on one side against the buffer with the dialysis membrane.

With a dialysis tube it is also possible to concentrate samples. The dialysis tubing is not to be put in a buffer tank, but in a hygroscopic material, for example, Sephadex G100. This draws the liquid and small molecules through the membrane wall. The material will be changed if it becomes wet.

Ultrafiltration and Diafiltration For rapid concentration of protein solutions ultrafiltration is suitable. Asymmetric membranes, having different pores size at the top and bottom and various exclusion limits, of cellulose, cellulose esters, polyethersulfone, or poly(vinylidene fluoride) (PVDF) have been developed. Ultrafiltration is not like dialysis, which is driven by a concentration gradient, but works according to the flow rate of the solvent through the ultrafiltration membrane. The salts (or other molecules with molecular weights significantly below the cutoff) are pressed together with the water through the membrane. For this purpose, pressure, vacuum, or centrifugation can be used. Ultrafiltration is usually performed in disposable containers of different sizes, depending on the sample volume. The sample volume is thereby reduced without a significant change in the salt concentration.

Diafiltration is based on the same principle, in which the sample is desalted in a similar way to dialysis. Here the volume of sample is kept constant by continuously feeding fresh dialysis buffer to the sample to replace the volume that is removed to the filtrate during ultrafiltration. The time needed for a particular diafiltration can be estimated from the volume of the filtrate. For a reduction in salt concentration by a factor of 10 the volume of the collected filtrate must be 2.3 times the initial volume of the sample, for a 100-fold reduction in salt concentration 4.6-times the initial sample volume must be collected. In a variant of the technique the sample may be diluted after the filtration volume reduction and ultrafiltered again. This is repeated until the desired salt concentration is reached.

Gel Chromatography During gel chromatography the substances present in the sample are separated according to their size. Salts thereby elute after the passage of about one column volume. Gel chromatography has some drawbacks that make it less favorable for large quantities and large volumes, as often occur at the beginning of a purification procedure:

- The sample volume shall not exceed 5% of the column volume, otherwise the separation between proteins and salts is no longer sufficient.
- Gel chromatography columns are easily overloaded with large amounts of protein, which leads to a mixture of protein and salt ranges.
- The separation is better at low flow rate, resulting in a relatively long analysis time.
- In gel chromatography the elution volume is at least triple that of the initial sample volume, which usually necessitates a concentration or a concentrating next step in a separation procedure. Today, for small sample volumes and small amounts of protein several smaller spin columns are on the market, which contain gel chromatography material and show excellent recoveries in protein desalting.

As with dialysis, concentration, desalting, or salt exchange should be performed at 4 °C. In addition, in ultrafiltration there is a risk that small amounts of protein in very dilute solutions (<20 µg ml^{-1}) adsorb to the membranes or the vessel walls. Small amounts of glycerol or detergents (e.g., Triton X-100) at concentrations below the critical micellar concentration (CMC, Section 1.8.1) added to the protein solution may sometimes help to minimize the losses caused by adsorption.

Chromatography, Chapter 10

Reversed-Phase Chromatography Especially for the late steps in a purification protocol a possibility for desalting of relatively hydrophilic proteins is reversed-phase chromatography. Additionally, this often results in a further separation of the applied sample. Using volatile solvents such as 0.1% trifluoroacetic acid in water, larger volumes of salt containing samples can be desalted, since the hydrophilic salts are not bound by reversed-phase material, while proteins are bound via their hydrophobic regions onto the column. With a gradient of an organic solvent (usually acetonitrile in 0.1% trifluoroacetic acid), the proteins can be eluted again. For the desalting of proteins a rather hydrophilic, large pore reversed-phase material should be used since hydrophobic, fine-pore material that is very suitable for peptide separations often gives very poor yields in the elution of proteins. The protein containing fractions are collected and subsequently the volatile organic solvent (e.g., acetonitrile) is evaporated in a vacuum centrifuge or with nitrogen. The disadvantages of desalting with reversed-phase chromatography are a possibility of protein denaturation, a generally poor recovery of hydrophobic proteins, and the high price of the column material. Proteins exhibit a special retention behavior when bound to reversed-phase material.

Above a certain concentration of the organic modifier the binding of the protein to the chromatographic material changes from a typical reversed phase mechanism to a normal phase behavior. This allows a desalting of proteins with an inverse gradient. The sample is applied in substantially pure organic solvent onto the reversed-phase column. The protein binds according to a normal phase mechanism, while salts and less hydrophobic impurities are eluted. The protein can now be eluted with a gradient to 0.1% trifluoroacetic acid in water.

Reversed Phase Chromatography, Section 10.4.2

For sample preparation/desalting of peptides for mass spectrometry disposable pipette tips containing a few microliters reversed-phase material are often used. When applied to such small columns, the peptides bind to the reversed-phase material and are concentrated there, where they can be washed and desalted at the same time using aqueous solutions (e.g., 0.1% trifluoroacetic acid). The bound proteins are eluted by small amounts of highly concentrated organic solvents. It is important that during the sample application no organic solvent is present in the sample, as this would interfere with the binding of the peptides on the reversed-phase material. If organic solvent from previous purification steps is present, its concentration must be reduced by evaporation or dilution to a maximum of 5%.

Immobilization on a Membrane An excellent possibility to prepare and also store proteins for further analyses (e.g., sequencing, amino acid analysis, mass spectrometry) is to transfer them to a chemically inert membrane (immobilization). The salt-containing protein solution can be applied onto a hydrophobic membrane (e.g., PVDF), in portions to a small area, and dried. At not too great amounts of salt the hydrophobic binding of proteins to these membranes is strong enough that even intensive washing with water does not remove the proteins from the membrane. Proteins can also be transferred to membranes by electroblotting. This is usually combined with a gel electrophoretic separation as the last step of the purification procedure. The immobilization on a membrane is only recommended for pure proteins as the last purification step of sample preparation for structure determination, since the recovery of the immobilized proteins from the membrane is extremely difficult.

Blotting, Section 11.7

1.7 Concentration

Especially after the first steps of a cleaning cycle one is often confronted with large volumes of dilute protein solutions. Furthermore, also, the ionic strength or pH values often do not fit the conditions needed for the next purification step. Therefore, methods must be used in order to both concentrate solutions and desalt or resalt. For recovery and protein stability reasons each protein purification should be performed in a minimum number of sub-steps. If possible, concentration and salt exchange/desalting should be combined in one step (see also Section 1.6).

Precipitation As discussed above (Section 1.4), precipitation is an effective way to concentrate a protein from its solution and at the same time largely free it of salts. After centrifugation (typically at low temperature), the supernatant is discarded and the pellet dissolved in the desired buffer for further separation or analyses steps. Problems arise in protein solutions containing detergents, since the precipitation generally gives only poor yields.

Dialysis and Ultrafiltration These are used in addition to desalting also for the concentration of samples (Section 1.6). Especially for the concentration of small volumes and smaller amounts of protein in the last purification steps prior to structure analysis, special microconcentrators are often used that exhibit, due to their design, only low adsorption losses.

Binding to the Chromatographic Material An elegant method by which to enrich and concentrate proteins and peptides also from large volumes is to bind them on reversed-phase material. The sample application buffer must be devoid of any substances that provide eluting conditions for the reversed-phase material (e.g., acetonitrile, alcohols, detergents). Alternatively, ion exchange materials can be used to enrich proteins and peptides. This is only successful if the ionic strength of the sample buffer is very low (<10 mM), which can usually be achieved through desalting or dilution.

1.8 Detergents and their Removal

Many proteins and enzymes in their natural environment are not surrounded by water but by hydrophobic lipid layers, the biological membranes. In cell disruption, these proteins remain associated with the membranes. Peripheral membrane proteins, which are only surface-associated to the membrane, usually can be solubilized by relatively mild treatments, without dissolving the membrane. High salt concentrations, extreme pH conditions, high concentrations of chelating agents (10 mM EDTA), or denaturing substances (8 M urea or 6 M guanidine hydrochloride) can be used. But as soon as the proteins are partially or almost completely embedded in the membrane (membrane anchor or integral membrane proteins), attempts to bring them into solution with normal aqueous buffers will fail. Today these proteins represent a major challenge for the purification and analysis techniques. If the lipids are removed from the membrane the proteins will, via their hydrophobic regions, form insoluble aggregates and precipitate or bind very strongly to all types of surfaces that exhibit a certain hydrophobic character. Such proteins can generally be brought and kept in solution only by using detergents in the solution.

1.8.1 Properties of Detergents

Detergents are amphiphilic molecules, that is, they have on the one hand a polar hydrophilic, possibly also ionic, moiety that is responsible for a good water solubility and on the other hand they have a nonpolar, lipophilic moiety capable of interacting with hydrophobic regions of a membrane protein; in this respect they approximately mimic the natural lipid environment of membrane proteins. An important property of detergents is their ability to form micelles. These are aggregates of different detergent molecules in which all the hydrophilic groups point towards the outside and all hydrophobic groups are directed inwardly (Figure 1.6). The size of micelles is dependent on the individual detergent (Table 1.4). Membrane proteins are incorporated as hydrophobic molecules in these micelles and can be brought into solution outside of a biological membrane and often retain their biological activity. The lowest concentration at which detergents can form micelles is called the critical micellar concentration (CMC). The CMC is different for different detergents (Table 1.4) and is a function of parameters such as temperature, ionic strength, pH, and the presence of di- and trivalent cations or organic solvents. The CMC of ionic detergents decreases significantly with higher ionic strength, and is independent of temperature. However, the CMC of non-ionic detergents is relatively independent of the salt concentration, but significantly increases with increasing temperature.

As can be seen from Table 1.4, several different detergents can be used in the purification and analysis of membrane proteins. Which detergent for a particular protein is optimal must be

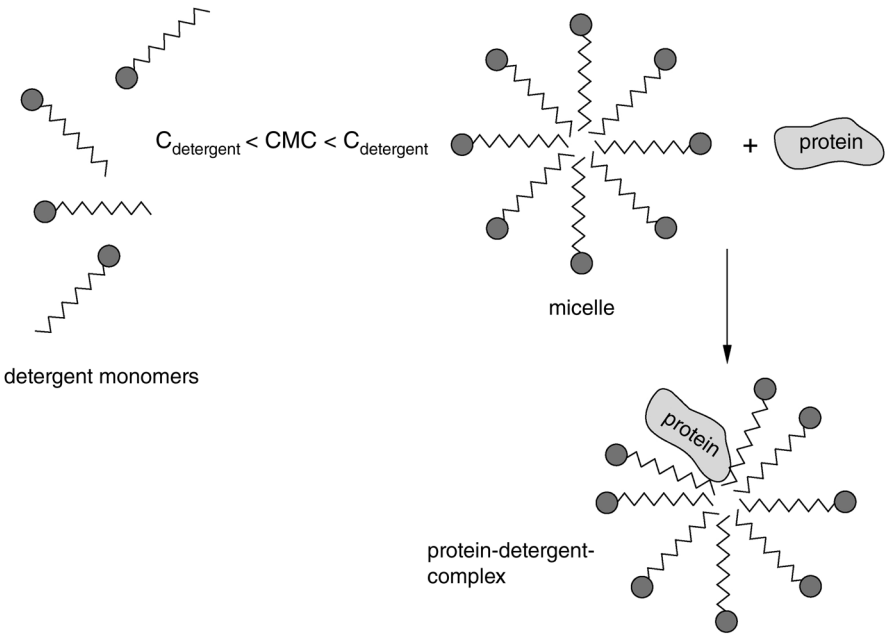


Figure 1.6 Formation of detergent micelles above the CMC and embedding of proteins into the micelles. In the micelle the hydrophilic polar structures face the aqueous phase and the hydrophobic parts are positioned inside. Membrane proteins can insert into such micelles and are kept in solution as a detergent–protein complex.

Table 1.4 Common detergents (*n*: aggregation number to form micelles).

	Name	M_r (monomer)	*n*	M_r (micelle) (kDa)	CMC (mM)	Dialyzable	Concentration of use	Comments
Ionic detergents	Sodium dodecyl sulfate (SDS)	288	60–100	>18	8	–	10 mg mg^{-1} protein	CMC is strongly dependent on ionic strength
	Desoxycholate	416	10	4	4 (in 50mM NaCl)	–	0.1–10 mg mg^{-1} membrane lipid	precipitates at low pH in presence of divalent cations, *n* increases with ionic strength
	Benzyldimethyl(*n*-hexadecyl) ammonium chloride (BAC-16)	396						cationic detergent, suitable for electrophoresis of membrane proteins
	Cetyltrimethylammonium bromide (CTAB)	365	170	62	1	–		
Non-ionic detergents	Triton X-100 poly(ethylene glycol)$_n$-octylphenyl ether	628	140	90	0.2	–	1–5 mM	Absorbs in UV (due to aromatic ring)
	Triton X-114	≈514			0.2	–	5 mM	phase separation at heating to over 22 °C
	Tween 80 (poly(oxyethylene)$_n$-sorbitan monooleate	≈1 310	58	76	0.02	–	>10 mg mg^{-1} membrane lipid	Polysorbates have been reported to be incompatible with alkalis, heavy metal salts, phenols, and tannic acid. They may reduce the activity of many preservatives
Zwitterionic detergents	Octylglucoside 1-Octyl-ß-D-glucopyranoside	292	30–100	8	15	+	20–30 mM	
	CHAPS 3-[(3-Cholamidopropyl)-dimethylammonio]-1-propanesulfonate	615	4–14	6	4	+	6–13 mM	Suitable for protein solubilization for isoelectric focusing/2D electrophoresis.
	Zwittergent 3-12 *N*-Dodecyl-*N*,*N*-dimethyl-3-ammonio-1-propanesulfonate	335	55	18	3.6	+	15–30 mM	

determined empirically for each individual case. Generally, detergent concentrations of 0.01–3% are used. The solubilization of proteins is often carried out at a detergent concentration close to the CMC.

Depending on the planned further purification steps, certain properties of individual detergents have to be taken into consideration:

- The high UV absorption of certain detergents (e.g., as in Triton or Nonidet due to the aromatic ring) can interfere with the detection of proteins, for example, in chromatographic analyses.
- Ionic detergents cannot be used in ion exchange chromatography, as they compete with the proteins for the ionic binding sites on the column material.
- Detergents often bind strongly via hydrophobic interactions on reversed-phase columns and also change the elution properties of the reversed phase materials.
- How easily can the detergent be removed from the protein (Section 1.8.2)?
- Especially, non-ionic detergents of the oxyethylene family (Triton, Tween, Nonidet) can easily contain oxidative impurities, particularly peroxides. Since these impurities can modify proteins, these detergent solutions should be as fresh as possible, be stored under nitrogen, and removed with a syringe from the vessel.

Whether a detergent has brought a certain protein in solution can be decided by centrifugation. A soluble protein (or its activity) has to be found in the supernatant after one hour of centrifugation at 100 000*g*. If no biological activity can be measured with various detergents, the buffer conditions should be changed or substances should be added that are known to stabilize the protein structure, for example, 20–50% glycerol, reducing agents such as dithiothreitol, chelating agents such as 1 mM EDTA, or protease inhibitors.

Ionic Detergents with cationic or anionic hydrophilic groups solubilize proteins well in its monomeric form and also solubilize membrane proteins. However, these detergents have the disadvantage of distorting and denaturing the structure of proteins – thus they almost always lose their biological activity. Probably the most widely applied detergent, sodium dodecyl sulfate (SDS), is used in the electrophoretic separation and molecular weight determination in polyacrylamide gels. The salt concentration significantly affects the CMC of SDS (salt free: 8 mM, in the presence of 0.5 M NaCl: 0.5 mM). The cationic detergent 16-BAC allows good separations of (membrane) proteins in an acidic environment. Thus, base-labile protein modifications remain intact (e.g., protein carboxymethyl esters) that are hydrolyzed in conventional purification methods in acidic or alkaline medium. 16-BAC is also successful as the first dimension of a two-dimensional gel electrophoretic separation of membrane proteins. The second dimension separation is a conventional SDS-gel electrophoresis.

SDS Gel Electrophoresis, Section 11.3.9

Nonionic Detergents These detergents affect less strongly protein–protein interactions and are generally less denaturing than ionic detergents. They are therefore particularly suitable for the isolation of intact protein assemblies, but in integral membrane proteins they do not prevent aggregation and furthermore, due to its low CMC, they are difficult to remove by dialysis. A special position is held by Triton X-114, which is soluble in water at 4 °C, but at temperatures above 20 °C forms water-insoluble micelles and thus shows a separation between the aqueous and detergent phase. Hydrophilic proteins remain in the aqueous phase, while integral membrane proteins are found in the detergent phase. This makes it possible to differentiate soluble proteins and membrane proteins.

Zwitterionic Detergents These detergents, which carry positively and negatively charged groups in their polar moiety, exhibit dissociating and denaturing properties between those of ionic and nonionic detergents. They can be applied in ion exchange chromatography or isoelectric focusing.

Ion Exchange Chromatography, Section 10.4.7

Isoelectric Focusing, Section 11.3.12, 12.4.7

1.8.2 Removal of Detergents

For subsequent analysis (amino acid analysis, amino acid sequence analysis, mass spectrometry, etc.) detergents, which are present in relatively high concentrations, interfere almost always

and usually must be removed. All processes for the removal of detergents should considered, however, that the solubilized hydrophobic proteins almost always need the presence of these substances for their solubility and activity. Otherwise, activity losses or adsorption losses to all types of surfaces will occur! The removal of detergents should therefore always be planned as the last step before the actual analysis so that no further sample manipulation takes place.

Dilution below the Critical Micelle Concentration Micelles of detergents with a high CMC (e.g., octylglucoside) can be dissolved into their monomeric state by diluting below their CMC and can then be easily removed by dialysis (caution: adsorption of detergent-free proteins to the dialysis membranes). Detergents with a low CMC (most of nonionic detergents) are virtually impossible to remove by dialysis and are best removed by special chromatographic columns (see below).

Extraction Different extraction methods are used to remove SDS. Besides the chloroform/methanol extraction, ion-pair extraction is especially commonly used. The dry sample is extracted with a solution of the ion-pair reagent in an organic solvent. Typical systems are acetone/triethylamine/acetic acid/water or heptane/tributylamine/acetic acid/butanol/water. Sufficient water (about 1%) must be present in order to form the relatively nonpolar alkyl ammonium SDS-ion pair which is extractable with organic solvents. Specimens that are present even in small volumes of aqueous solution can be freed of SDS; only the water is omitted from the extraction solutions. This method can remove in an extraction step up to 95% of SDS. The protein is recovered as a precipitate by centrifugation and residual SDS can then be removed by washing with heptane or acetone. Salts may interfere with the removal of SDS and should be removed prior to extraction. The recovery rates for proteins are often 80% and higher.

Ionic Retardation Chromatographic materials from polymerized acrylic acid that includes a strong ion exchange resin (quaternary ammonium groups on a polystyrene–divinylbenzene matrix) are available. SDS–protein complexes, which are chromatographed on these materials, lose almost all of SDS, and the protein is eluted with high yields of 80–90%. These columns are designed as single use columns, since the bound SDS practically cannot be removed. Also with ionic retardation, salts may interfere with the SDS binding to the column material and should be removed beforehand. This can be done by gel filtration or by simply applying a small layer of size exclusion gels at the top of ion retardation column, resulting in delay of the proteins and hence a separation from the buffer salts.

Gel Filtration in the presence of Organic Solvents In particular relatively weak hydrophobic proteins can sometimes be kept in solution in a buffer containing organic solvent (e.g., 20% acetonitrile), and thus separated from salts and detergents by a gel permeation chromatography. More hydrophobic proteins must be kept in solution by suitable solvent mixtures having a high content of organic acids, such as formic acid/propanol/water.

Size Exclusion Chromatography, Section 22.7

Removal of Detergents by Reversed-Phase HPLC Membrane proteins can be freed from detergents and salts also by using reversed-phase column chromatography. Short, relatively hydrophilic columns (RP-C4) with a low surface area must be used to achieve satisfactory recoveries. The chromatographic conditions are generally selected in such a way that a compromise between recovery and the quality of separation is obtained. Generally, short, steep gradients using volatile solvent systems with high levels of organic acids (see above) should be used. These methods are mainly used when the primary objective is the structure determination of membrane proteins, since the biological activity is almost always destroyed by the high proportion of organic solvent and acid. For desalting, removal of detergents and hydrophobic components, reversed-phase chromatography has been applied successfully also with an inverse gradient (see Section 1.6, reversed-phase HPLC).

Reversed Phase Chromatography, Section 10.4.2

Special Chromatographic Support Materials for the Separation of Detergents Removal of detergents by specific, commercially offered materials is based in principle on a hydrophilic column material with such a small pore volume that proteins with a molecular weight of above 10 000 Da cannot enter the interior of the stationary phase and elute in the void volume. The relatively small detergent molecules, however, reach the interior of the column material, where

they interact with hydrophobic binding sites and are bound there. Binding capacities for the different detergents are from 50 to 100 mg detergent per ml of column material. Protein concentrations should be above 50 µg ml^{-1}, because otherwise significant losses due to nonspecific adsorption to the column material can be found.

Blotting, Section 11.7

Blotting onto Chemically Inert Membranes The easiest way to purify a hydrophobic protein and analyze it is to blot it directly from a detergent containing polyacrylamide gel electrophoresis (SDS-gel) onto a chemically inert membrane. The immobilized protein can then be directly subjected to further analysis, for example, amino acid analysis, sequence analysis, immunological methods, or even mass spectrometry.

1.9 Sample Preparation for Proteome Analysis

The aim of a meaningful proteome analysis is to determine the quantitative relationships of all the proteins in a proteome sample. The ratio of each individual protein in different proteomic states must not change due to the influence of sampling, separations, or analysis procedures. This has the consequence that expensive multistep purification processes, such as those used in classical protein purification applications, can no longer be applied because considerable losses in the individual steps may occur and, even more importantly, they are different and unpredictable for each individual protein. The sample preparation for proteome analysis is very dependent on the biological question and the starting material and therefore cannot generally be treated. But, in principle, all the proteins in a proteome should be brought into solution in a form in which they are directly suitable for separation, for example, for a two-dimensional gel electrophoresis or HPLC-MS analysis.

Proteome Analysis, Section 39.3

Further Reading

Bollag D.M. and Edelstein S.J. (1991) *Protein Methods*, Wiley-Liss, New York.

Deutscher, M.P. (ed.) (1990) *Guide to Protein Purification*, Methods in Enzymology, vol. 182, Academic Press.

Rickwood D. (ed.) (1991) *Centrifugation: a Practical Approach*, 2nd edn, IRL Press, Oxford.

Seddon M.A., Curnow P., and Booth P.J. (2004) Membrane proteins, lipids and detergents: not just a soap opera. *Biochim. Biophys. Acta.*, **1666**, 105–117.

Protein determination

2

Lutz Fischer and Timo Stressler

University of Hohenheim, Institute of Food Science and Biotechnology, Chair of Biotechnology and Enzyme Science (150b), Garbenstraße 25, 70593 Stuttgart, Germany

The determination of the protein content of aqueous solutions is a frequent task in biochemical and biological laboratories. Therefore, it is important to know the principles and especially the advantages and disadvantages of the main methods. Furthermore, one must be aware of their method-related uncertainties and sources of error. Objective methods for the quantification of proteins are quantitative amino acid analysis or weight determination of the pure protein as a solid. The latter is possible after sample preparation by vacuum freeze-drying or heat-drying at 104–106 °C for 4–6 h. However, both methods are too time-consuming and laborious for the routine determination of the protein content of a solution. It is also quite sufficient for many issues to know the approximate contents compared to a defined standard, such as serum albumin. This applies, for example, to protein determinations made during protein purification or the immobilization of proteins.

Analysis of Amino Acids, Chapter 13

The proteins are irreversibly denatured by the commonly used colorimetric methods due to the color reaction in the strongly basic or acidic environment. By contrast, spectroscopic methods preserve the biological function and the protein can be reused. Unfortunately, spectroscopic methods have lower sensitivities than colorimetric methods. Both methods can be greatly disturbed and affected by the presence of other substances. Therefore, the presence of salts, sugars, nucleic acids, or detergents must be taken into account when deciding which of the following methods for quantification presented are applicable or not.

Complexity and Individuality of the Proteins Proteins are composed of up to 20 different amino acids. They are present in different, individual ratios. The total number of amino acids also varies from protein to protein. In addition, prosthetic groups and sugar structures increase the complexity. Neighboring group effects in proteins also often influence the chemical or physical behavior of functional groups. There is no color reaction or spectroscopic method that can respond in the same way to *all* the properties of a protein. Therefore, the reaction with a dye or the absorption or emission at a particular wavelength relates to each of a few functionalities or groups of the protein. If they are accidentally present in a protein or protein mixture, an erroneously higher protein content will be measured and vice versa.

Some color reactions are based on the complex forming or reducing property of the peptide bond. Therefore, the number of peptide bonds in a protein is a comparatively good means for the quantification of proteins. This is because such reactions are less dependent on the individual amino acid composition. Other methods are based more on the side chain properties of the amino acids – ionic, hydrophilic, hydrophobic, or aromatic – and, therefore, they are more subjective. None of the methods presented in this chapter is "objective" and "completely accurate"! Nevertheless, they have been proven in practice. However, one must know their limitations and take into account the fact that each method is more or less of a compromise.

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

Quantification of Complex Protein Mixtures or Pure Protein Solutions? The scope – the aim of the analysis – is in deciding:

- which of the protein determination methods to choose,
- from which basic data (references) the quantification should occur.

If, for example, the protein content of a complex protein mixture, such as the total protein content of cell material, has to be determined, the methods chosen should detect relatively objective functions in proteins, such as peptide bonds. The analysis should be realized based on well-characterized analytically "mean proteins," standard proteins such as bovine serum albumin or chymotrypsin. Using a standard protein, a calibration curve can be prepared or data from the literature can be used.

By contrast, if the sample is a less complex protein solution, such as an already purified and enriched protein fraction, and a specific protein needs to be quantified specifically, all methods can be taken into account in principle. It is crucial that the target protein is available as a standard protein (calibration curve) or literature data of the target protein are known. In all cases, it should be noted that identical conditions (solutions, pH, temperature, incubation times, etc.) must be satisfied in the test solution and the reference.

Non-protein Components in Protein Solutions The significance of non-protein compounds in sample solutions has already been mentioned. They can significantly manipulate and falsify the dye formation or absorption measurement. However, proteins are macromolecules with a molecular weight of usually over 10 000 Da and the disturbing non-protein molecules are predominantly of low molecular weight, of less than 1000 Da. Therefore, they can be separated from each other by simple dialysis, acid precipitation, gel exclusion chromatography, or ultrafiltration. For the two latter methods, laboratory equipment, specialized for small volumes (0.5–5 ml), is available from various suppliers. Because some non-protein substances interfere with *one* assay, but not with the *other* assay, it is often sufficient to switch to another method. This must be decided on a case-by-case basis. An indication of this is given in Table 2.1.

Table 2.1 Selection of some substances that interfere with the different color assays and UV methods.

Method for protein determination	Interfering substances
Biuret assay	ammonium sulfate glucose sulfhydryl compounds sodium phosphate
Lowry assay (modified according to Hartree)	EDTA guanidine-HCl Triton X-100 SDS Brij 35 >0.1 M TRIS ammonium sulfate 1 M sodium acetate 1 M sodium phosphate
Bicinchoninic acid assay	EDTA > 10 mM sucrose or glucose 1.0 M glycine >5% ammonium sulfate 2 M sodium acetate 1 M sodium phosphate
Bradford assay	>0.5% Triton X-100 >0.1% SDS sodium deoxycholate
UV methods	pigments phenolic compounds organic cofactors

Additional Methods, not described in detail In addition to the methods described in this chapter, the titrimetric determination of nitrogen by the Kjeldahl method and the ninhydrin assay can be used for the quantification of proteins after acid, thermal degradation of the proteins. Neither method is addressed in detail as they require a great deal of effort, but are discussed here briefly for the sake of completeness.

Using the Kjeldahl method, organic nitrogen compounds are oxidized to CO_2 and H_2O and create an equivalent amount of NH_3. Defined conditions are used for the heating with concentrated sulfuric acid and a catalyst (heavy metals, selenium). The NH_3 obtained is bound by H_2SO_4 as $(NH_4)_2SO_4$ (wet ashing). After the addition of NaOH, ammonia is released, transferred to a distillation apparatus and quantified by titration. The nitrogen content of proteins is approximately 16%. Therefore, multiplying the N-content determined by 6.25 can recalculate the amount of protein. Obviously, the non-protein nitrogen must be have been previously removed.

The color assay with ninhydrin is used as a detection method for free amino groups. Therefore, the protein must, firstly, be hydrolyzed into its free amino acids. Exemplarily, this is realized by boiling in 6% sulfuric acid at 100 °C (12–15 h) in fused glass vessels in the absence of oxygen. The ninhydrin reagent is added to the protein hydrolyzate and the resulting purple–blue solution is measured spectrophotometrically at a wavelength of 570 nm. In most cases, L-leucine is used as a standard to generate the calibration curve. However, the color intensities, resulting from the different amino acids of the protein, are not identical. This is one of several sources of error in the ninhydrin method.

Ninhydrin Assay, Section 13.3.1

2.1 Quantitative Determination by Staining Tests

Protein samples often consist of a complex mixture of different proteins. The quantitative determination of the protein content of such crude protein solutions is usually based on the color reactions of functional groups of proteins with dye-forming reagents. The intensity of the dye correlates directly with the concentration of the groups reacting and can be accurately measured with a spectrophotometer. The basics of spectroscopy (Lambert–Beer law, etc.) and the appropriate equipment are described in detail in Chapter 7. There are several variants for the four following staining methods, which are described in the literature. However, they are all based on the same principles.

Multiple determinations should be performed in all cases. Triplicate measurements are usually realized and the mean value calculated. The samples are generally measured at the same wavelength against a so-called blank approach, which consists of the same ingredients and volumes as the respective color assay but the protein solution is replaced by distilled water.

Spectral Absorption Coefficients Each staining method can only be used in a certain concentration range. Within this range, a constant dependency of the absorption measured on the protein concentration results (at a defined wavelength). The spectral absorption coefficient is determined graphically as the slope in the plot of absorbance versus concentration (abscissa). By default, the absorbance value is related to the path length of the cuvette (in cm) and the concentration of the dissolved protein value in micrograms per milliliter. Alternatively, with a known molecular weight of the protein, the concentration unit mole of dissolved protein per milliliter may be used. Then, a molar spectral absorption coefficient results (formerly molar extinction coefficient) with the units: 1/(moles of dissolved protein per liter) per cm or liters per mole of dissolved protein per cm.

Physical Principles of Spectroscopy, Section 7.1

The requirements for the staining methods presented (protein concentration ranges, sample volumes) and the approximate resulting spectral absorption coefficient (ml final volume per microgram of protein dissolved per cm), with bovine serum albumin as a standard, are shown in Table 2.2 as an overview. Approximate values are presented in the table because spectral absorption coefficients between 2.3 and 3.2 ml final volume per microgram of protein in solution per cm can be found under apparently identical conditions in the literature (e.g., solely for the biuret-assay)! This is caused by the complexity of influencing factors such as the purity of the chemicals and the water used.

It is very important for all staining methods to specify what the volume (ml) stands for. Several solutions have to be combined in different volumes with the protein sample, depending on the method. The volume specified should always stand for the *final volume* of the test approach after performing the assay and *not* the volume of the protein solution used.

Relative Deviations of the Staining Methods Ideally, non-proteinogenic impairment of the assays can be excluded and, apart from a few exceptions, appear under the determination methods presented for one and the same protein, deviating between at least 5% and 20%. The difference is even more dramatic for the quantification of crude protein solutions. It is extremely important

Table 2.2 Overview of the most common staining methods for protein determination.

Method	Approx. sample volume required (ml)	Limit of detection (µg-protein ml^{-1})	Spectral absorption coefficient[a] (ml final volume per µg dissolved protein per cm)
Biuret assay	1	1–10	$2.3 \times 10^{-4}\ A_{550}$
Lowry assay (modified according to Hartree)	1	0.1–1	$1.7 \times 10^{-2}\ A_{650}$
Bicinchoninic acid assay	0.1	0.1–1	$1.5 \times 10^{-2}\ A_{562}$
Bradford assay	0.1	0.05–0.5	$4.0 \times 10^{-2}\ A_{595}$

a) With the standard protein bovine serum albumin.

when reporting the specific activities of enzymes, antibodies, or lectins, expressed as biological activity per mg of protein, not only to state under which test conditions (e.g., substrate, pH, temperature) the activity was determined but also *which method* was used for the protein determination.

2.1.1 Biuret Assay

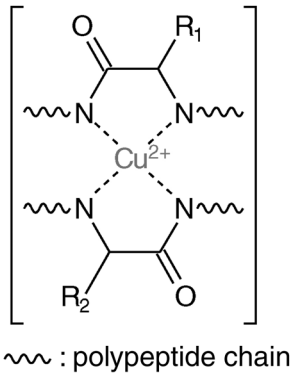


Figure 2.1 The colored protein Cu^{2+} complex that occurs in the biuret reaction.

This protein determination method is based on a color reaction with dissolved biuret (carbamylurea) and copper sulfate in an alkaline, aqueous environment (biuret reaction). The result is a red-violet colored complex between the Cu^{2+} ions and two molecules of biuret. The reaction is typical of compounds with at least two CO-NH groups (peptide bonds) and can, therefore, be used for the colorimetric detection of peptides and proteins (Figure 2.1). If tyrosine residues are present, they also contribute significantly to the dye formation by the complexation of copper ions. Thus, the detection is mainly oriented objectively on the peptide bonds and subjectively on the tyrosine residues of proteins. The spectral absorption coefficient given in Table 2.2 was determined at 550 nm. Otherwise, the color intensity can also be measured at 540 nm. Both wavelengths are close to the absorption maximum of the color complex. The absorption maximum varies slightly from protein to protein.

The biuret assay is the least sensitive of the color assays (Table 2.2). The protein sample or standard sample is mixed with four parts of biuret reagent and allowed to stand for 20 min at room temperature. Then, the color intensity is directly measured in a spectrophotometer. Ammonium and weak reducing and strong oxidizing agents act especially in a disturbing way (Table 2.1). However, small amounts of sodium dodecyl sulfate (SDS) or other detergents are tolerable. If the solution has to be diluted due to the high absorption, this must done with the sample solution used and not with the final solution after color formation. The color formation reaction must be repeated. This ensures that the required amount of copper ions are present, due to the concentration-dependent equilibrium, which is necessary for complete saturation settings of the complex-forming groups.

2.1.2 Lowry Assay

Lowry *et al*. published a method for the quantitative analysis of proteins in 1951. The method is a combination of the biuret reaction and the Folin–Ciocalteau phenol reagent and is referred to as the Lowry assay. The copper–protein complex mentioned above is formed in alkaline solution. This supports the reduction of molybdate or tungstate, which are used in the form of their heterogenic polyphosphoric acid (Folin–Ciocalteau phenol reagent), by primarily tyrosine, tryptophan, and, to a lesser extent, cysteine, cystine, and histidine of the protein. Here, presumably, Cu^{2+} of the copper–protein complex is reduced to Cu^{+}, which subsequently reacts with the Folin–Ciocalteau phenol reagent. Due to the additional color reaction, the sensitivity increases enormously compared to the pure Biuret assay. The resulting deep blue color is measured at a wavelength of 750, 650, or 540 nm.

Various modifications for the Lowry assay are described in the literature. The aim was mostly to improve the relatively high breakdown susceptibility of the Lowry method. The data presented in Tables 2.1 and 2.2 were obtained using a published version of *Hartree* (1972). The modified method extends, by the same sensitivity, the linear range, compared to the conventional Lowry assays, by 30–40% to about 0.1–1.0 mg ml^{-1} (Table 2.1). The method shows no problems with dropout salts and used only three stock solutions, which also have better storage stability, instead of the five stock solutions of the original Lowry assay. In this variant, three reagents (parts A : B : C = 0.9 : 0.1 : 3.0) are added successively to one amount of protein sample (1.0): A (carbonate/NaOH solution), B (alkaline $CuSO_4$ solution), and C (diluted Folin–Ciocalteau reagent). After the addition of A and C, the mixture is heated to 50 °C for 10 min each time. Overall, the Lowry assay according to Hartree takes about 30 min. Any necessary dilutions must, as already explained for the biuret assay, be performed with the protein solution.

The Lowry method is affected by a wide range of non-proteinogenic substances (Table 2.1). In particular, the usual additives for enzyme purification, such as EDTA, ammonium sulfate, or Triton X-100, are not compatible with the Lowry assay. Compared with the biuret assay, subjective criteria contribute more intensely to dye formation – in particular the individual rates, depending on the protein, of tyrosine, tryptophan, cysteine, cystine, and histidine. Again, the staining is relatively unstable. The measurement of the samples should be carried within 60 min of the last reaction step.

2.1.3 Bicinchoninic Acid Assay (BCA Assay)

Smith and colleagues published a highly regarded alternative to the Lowry assay in 1985 that combines the biuret assay with bicinchoninic acid (BCA) as the detection system.

Hitherto, BCA had been used for the detection of other copper-reducing compounds, such as glucose or uric acid. Twenty parts of a freshly prepared bicinchoninic acid/copper sulfate solution is added to one part of sample and incubated for 30 min at 37 °C.

Similar to the Lowry assay, the method is based on the reduction of Cu^{2+} to Cu^{+}. The BCA forms a color complex specifically with Cu^{+} (Figure 2.2). This allows a sensitive, colorimetric detection of proteins at a wavelength of 562 nm (the absorption maximum of the complex). Comparisons with the Lowry assay showed that cysteine, cystine, tyrosine, tryptophan, and the peptide bond reduce Cu^{2+} to Cu^{+} and, therefore, allow the color formation with BCA. The intensity of the color formation and the redox behavior of the groups involved depend on, among other things, the temperature. Thus, the BCA assay can be varied between different *temperatures* to obtain the sensitivity desired.

The BCA and Lowry assays are in good agreement for the determination of the concentrations of standard proteins, such as bovine serum albumin, chymotrypsin, or immunoglobulin G. Significant deviations of almost 100% were determined with avidin, a glycoprotein from chicken egg-white. The mechanism of the BCA assay is similar in principle to that of the Lowry assay. However, in no cases are they equivalent. The advantages of the BCA assay over the Lowry assay are the simpler implementation, the ability to influence the sensitivity, and the good stability over time of the color complex formed. The disadvantage is the higher price of the assay, due to the high price of the sodium salt of bicinchoninic acid. The sensitivity of the BCA assay is in the range of the Lowry assay modified by Hartree (Table 2.2). The breakdown

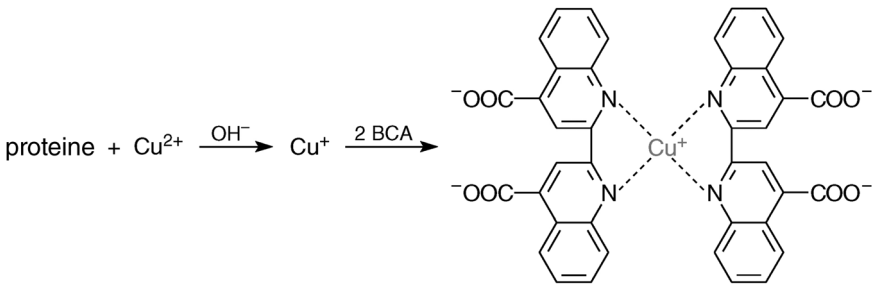


Figure 2.2 The bicinchoninic acid assay: a combination of the biuret reaction with the selective bicinchoninic acid complexation of Cu^{+}.

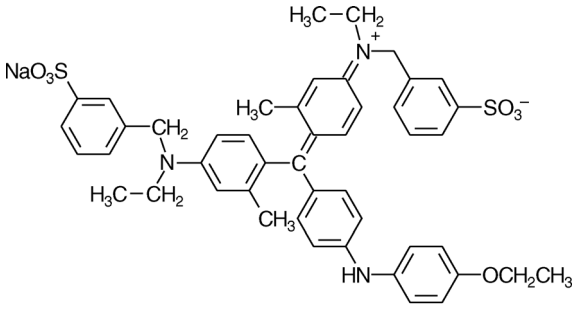


Figure 2.3 Coomassie Brilliant Blue G250 (as sulfonate), the reagent of the Bradford assay.

susceptibility of the BCA assay is also quite high. In addition to the substances listed in Table 2.2, further substances, such as small amounts of ascorbic acid, dithiothreitol, or glutathione, complexing and reducing compounds, interfere with the assay.

2.1.4 Bradford Assay

In contrast to the dyeing methods described so far, no copper ions are involved in this assay. It is named after M.M. Bradford and was published in 1976. The focus is on blue acid dyes, which are called Coomassie Brilliant Blue. In many cases, Coomassie Brilliant Blue G 250 (Figure 2.3) is used. The absorption maximum of Coomassie Brilliant Blue G 250 shifts from 465 to 595 nm in the presence of proteins and in an acidic environment. The reason for this is probably the stabilization of the dye in its unprotonated, anionic sulfonate form by complex formation between the dye and protein. The dye binds *fairly non-specifically* to cationic and nonpolar, hydrophobic side chains of proteins. The interactions with arginine, and less so with lysine, histidine, tryptophan, tyrosine, and phenylalanine, are most important.

Protein Detection in Electrophoresis Gel, Section 11.3.3

The Bradford assay is also used for staining proteins in electrophoresis gels. It is approximately a factor of two more sensitive than either the Lowry or BCA assay (Table 2.2) and is, thus, the most sensitive quantitative staining assay. It is also the simplest assay, because the stock solution, consisting of dye, ethanol, and phosphoric acid, is added to the sample solution in a ratio of 20-to-50 : 1 and, after 10 min at room temperature, the measurement of the absorbance at 595 nm can be started. Another advantage is that several substances that interfere with the Lowry or BCA assay do not affect the result of the Bradford assay (Table 2.1); this is especially the case concerning the tolerance to reducing agents! On the other hand, all substances that affect the absorption maximum of Coomassie Brilliant Blue are disturbing. This is sometimes difficult to estimate beforehand due to the lack of specificity of the interactions. The biggest disadvantage of the Bradford assay is that equal amounts of different standard proteins can cause significant differences in their resulting absorption coefficient. Thus, the *subjectivity* of this color assay is considerable and is bigger than that of the three other more complex staining methods.

2.2 Spectroscopic Methods

Spectroscopic Bases and Measurement Techniques, Section 7.1

Spectroscopic methods are less sensitive and require higher concentrations of protein than colorimetric methods. These spectroscopic methods should be used with purer or high-purity protein solutions. The spectral absorption or emission properties of the proteins at a defined wavelength in an optical pathway are measured. Therefore, the protein solution (sample solution) is placed in a quartz cuvette (optical pathway in the cell: usually 1 cm). The spectrophotometer is previously calibrated with pure, protein-free solvent in the same quartz cuvette and set at zero as a reference.

Subsequently, the value of the sample solution measured will result, either based on literature tables or the calibration curve, in the corresponding protein concentration in mg ml^{-1}. The latter

Table 2.3 Overview of the most common spectroscopic protein determination methods.

Method	Protein component on which the determination is essentially based	Limit of detection (µg-protein ml^{-1})	Dependence on protein composition	Susceptibility to interference
Photometry:				
A_{280}	Tryptophan, tyrosine	20–3,000	Strong	Low
A_{205}	Peptide bonds	1–100	Weak	High
Fluorimetry:				
excitation$_{280}$ emission$_{320–350}$	Tryptophan (tyrosine)	5–50	Strong	Low

The absorbance value determined (sample or standard) should not exceed 1.0. At a value greater than 1.0, the *linear dependency* of spectral absorption on the concentration is *no* longer given. The emission value for fluorescence measurements should not exceed 0.5. If necessary, the sample solution must be diluted and the dilution factor has to be taken into consideration when determining the concentration.

is recommended due to the interference effects of buffer substances, pH values, inaccuracies of devices, and so on. Ideally, a calibration curve is obtained with the pure protein of interest as a standard.

An overview of the detailed spectroscopic methods discussed below is given in Table 2.3.

2.2.1 Measurements in the UV Range

Absorption Measurement at 280 nm (A_{280}) Warburg and Christian measured the protein concentration of cell extract solutions of different purification degrees at a wavelength of 280 nm (A_{280}) in the early 1940s. The aromatic amino acids, tryptophan and tyrosine, and to a lesser extent phenylalanine, absorb at this wavelength (Table 2.4). Since larger amounts of nucleic acids and nucleotides are present in the protein solutions – this is generally the case after digestion of cells – the values measured at A_{280} had to be corrected. This is because the nucleic acid bases also absorb at A_{280}. Thus, Warburg and Christian determined a second value at 260 nm (A_{260}), which was correlated with A_{280} according to the following formula:

$$\text{Protein concentration}\left(\text{mg ml}^{-1}\right) = (1.55A_{280}) - (0.76A_{260}) \qquad (2.1)$$

This relationship can be used up to 20% (w/v) nucleic acids in solution or an A_{280}/A_{260} ratio of less than 0.6. The A_{280} measurement alone is sufficient in protein solutions with only a low content of nucleic acids.

In accordance with the molar spectral absorption coefficient (ε) (Table 2.4), the A_{280} method is based essentially on tryptophan, which has an absorption maximum at 279 nm. The two other aromatic amino acids contribute relatively less to the A_{280} value. Since the content of aromatic amino acids can vary from protein to protein, the corresponding A_{280} values also vary. At a concentration range of 10 mg ml^{-1} ($A^{1\%}$), the A_{280} value of most proteins is between 0.4 and 1.5. However, there are extreme exceptions where $A^{1\%}$ is 0.0 (parvalbumin) or 2.65 (lysozyme). An ideal standard protein should have the same level of aromatic amino acids as the protein measured, or should be identical with it. Unfortunately, this is extremely rarely realizable in practice.

The A_{280} method can be used for protein concentrations between 20 and 3000 µg ml^{-1}. It is an easy and fast method and is a lot less disturbed by parallel absorption of non-protein substances

Table 2.4 Molar spectral absorption coefficient (ε) at 280 nm and absorption maxima of aromatic amino acids.[a)]

Amino acid	$\varepsilon \times 10^{-3}$ (l mol^{-1} cm^{-1}) at 280 nm	Absorption maxima (nm)
Tryptophan	5.559	219, 279
Tyrosine	1.197	193, 222, 275
Phenylalanine	0.0007	188, 206, 257

a) For aqueous solutions at pH 7.1.

Table 2.5 Maximum concentration of disturbing additives allowed by the A_{205} and A_{280} method. The additives are often used in protein chemistry.[a)]

Additive	A_{205} method	A_{280} method
Ammonium sulfate	9% (w/v)	>50% (w/v)
Brij 35	1% (v/v)	1% (v/v)
Dithiothreitol (DTT)	0.1 mM	3 mM
Ethylenediaminetetraacetic acid (EDTA)	0.2 mM	30 mM
Glycerol	5% (v/v)	40% (v/v)
Urea	<0.1 M	>1 M
KCl	50 mM	100 mM
2-Mercaptoethanol	<10 mM	10 mM
NaCl	0.6 M	>1 M
NaOH	25 mM	>1 M
Phosphate buffer	50 mM	1 M
Sucrose	0.5 M	2 M
SDS (sodium dodecyl sulfate)	0.1% (w/v)	0.1% (w/v)
Trichloroacetic acid (TCA)	<1% (w/v)	10% (w/v)
TRIS buffer	40 mM	0.5 M
Triton X-100	<0.01% (v/v)	0.02% (v/v)

a) According to: Stoscheck, C.M. (1990) Quantitation of protein. *Methods Enzymol.* **182**, 50–68.

than the method described below. The latter is also a common method for protein determination using UV. This method measures the absorption in the lower UV region at 205 nm (A_{205}) and is significantly more sensitive (Table 2.5).

Absorbance Measurement at 205 nm (A_{205}) The A_{205} method was published by Goldfarb, Saidel and Mosovich in the early 1950s and supplies reliable values in a concentration range of 1–100 µg ml^{-1}. This method is based on the absorption properties of the *peptide bonds* and is, therefore, less dependent on the composition of the proteins. Unfortunately, in addition to the interference of various buffer substances (Table 2.5.), the absorption maxima of several amino acids (e.g., histidine 211 nm, phenylalanine 206 nm, tyrosine 193 and 222 nm, and tryptophan 219 nm) also interfere. In general, all non-protein molecules that contain C═C or C═O double bonds are trouble for the A_{205} method. A very clean quartz cuvette, a relatively new deuterium lamp in the spectrophotometer, and a minimum concentration of suitable buffers (Table 2.5) should be used for the realization of A_{205}.

When the influence of buffer and non-protein substances is excluded, the $A^{1\%}_{205}$ value of almost all proteins is within the range 28.5–33. If the $A^{1\%}_{205}$ value of the desired protein is not known, the protein concentration can be calculated using the following formula:

$$\text{Protein concentration}\left(\text{mg ml}^{-1}\right) = \frac{(31A_{205})}{X} \tag{2.2}$$

In this case, the A_{205} value has the unit mg ml^{-1} cm^{-1} and X is the thickness of the quartz cuvette (in cm), which is usually 1 cm. The accuracy of the method is about ±10%.

Scopes corrected the A_{205} method in 1974, for the content of tryptophan and tyrosine. Therefore, he determined the absorbance values both at A_{280} and at A_{205} and used them in the following formula:

$$\text{Protein concentration}\left(\text{mg ml}^{-1}\right) = 27.0 + 120\left(\frac{A_{280}}{A_{205}}\right) \tag{2.3}$$

In this case, the content of protein predicted in a sample had a maximum deviation of only 2%.

2.2.2 Fluorescence Method

Fluorescence Spectroscopy, Section 7.3

In this method, the intrinsic fluorescence (primary fluorescence) of the aromatic amino acids in proteins, mainly of tryptophan, is measured in a spectrofluorometer after appropriate excitation. With this method, a calibration curve must be created under identical conditions using the *protein of interest* as the standard protein. If this is not possible, the fluorescence method is only suitable for a *qualitative detection* of proteins in a solution. This is because the Lambert–Beer law, in contrast to the photometry, is not directly correlated.

Excitation of the fluorescent groups in the proteins generally takes place at 280 nm and the determination of the emission values is in the range 320–350 nm. The latter depends on the particular protein. The exact emission wavelength is determined by recording the emission spectrum after excitation at 280 nm. The wavelength at which the protein has its emission maximum will be set as a fixed value. The method is applicable in the concentration range 5–50 µg ml^{-1}, which is similar to that of the A_{205} method.

The fluorescence properties of the aromatic amino acids are shown in Table 2.6. Phenylalanine is not significantly detectable in the presence of the other two amino acids. One might conclude that tryptophan and tyrosine provide a nearly equal contribution for quantification. However, this is often not the case, since the fluorescence of the tyrosine residue is easily quenched. This requires only its ionization (pH dependent) or the close proximity of an amino or carboxy group or a tryptophan residue. This is often the case. Therefore, the tryptophan content of the protein is similar to that from the A_{280} method, which is critical for the fluorescence method, although tryptophan can also be quenched by neighboring acidic groups. An additional significant contribution to the quantum yield of a protein can come from fluorescent prosthetic groups.

In comparison to photometric methods, the fluorescence emission of proteins is linear over a smaller range. Furthermore, it is strongly dependent on the pH and the polarity of the solvent. For this method, disturbances by non-protein molecules need not to be further addressed. This is because the application can only be made in a very clean solution.

Table 2.6 Fluorescence properties of aromatic amino acids.[a)]

Amino acid	Excitation wavelength (nm)	Emission wavelength (nm)	Quantum yield[b)]
Tryptophan	285	360	0.20
Tyrosine	275	310	0.21
Phenylalanine	260	283	0.04

a) For aqueous solutions at pH 7.0 and 25 °C.
b) The quantum yield is the ratio of emitted to absorbed photons (see Sections 7.1 and 7.5).

2.3 Radioactive Labeling of Peptides and Proteins

In some areas of biochemistry, for example, for certain binding studies or radioimmunoassays, peptides or proteins are radiolabeled *in vitro* and then quantified according to the radioactivity. The high selectivity and sensitivity are the big advantages of this method. Thus, using a scintillation counter, a device for quantifying radioactivity, concentrations down to 10^{-15} mol l^{-1} of a peptide or protein, quadruple-labeled with ^{125}I, can be detected. Therefore, the limit of detection is several orders of dimensions below that of spectral methods.

The peptide or protein for radioactive labeling must be *purified* before the real investigations can begin. It is radioactively labeled using special chemical reagents and the average radioactivity per mole is determined (calibration). The labeled biomolecule is then ready-to-use and can be detected and quantified in various samples.

Peptides are mainly taken used for radioactive labeling. The methods are also suitable for proteins. However, the fact that the tertiary structure is affected and side reactions can take place on sensitive amino acids must be taken into account; either occurrence can easily destroy the biological activity.

Table 2.7 Overview of methods often used for the radiolabeling of peptides or proteins.

Amino acid residue	Established marker	Method
Histidine, tyrosine	^{125}I, ^{131}I	Chloramine T, Iodobeads
Tyrosine	^{125}I, ^{131}I	Lactoperoxidase
Lysine, N-terminus	^{125}I, ^{131}I	Bolton–Hunter reagent
Lysine, *N*-terminus	^{14}C, ^{3}H	(a) Anhydride (b) Aldehyde, borohydride
Cysteine	^{14}C, ^{3}H	Iodoacetic acid
Tyrosine	^{3}H	Reduction of ^{127}I
Each residue	^{14}C, ^{3}H	Peptide synthesis
Sugar moiety	^{3}H	Periodate, borohydride

Chemical Modifications and Methods, Chapter 6

Since only the target peptide is radiolabeled in a particular sample, the purification steps performed previously explain the selectivity of this method. An overview of the most common labeling strategies and markers is given in Table 2.7 and Figure 2.4.

The choice of the radioisotope depends largely on which amino acid residues are to be modified, which half-lives are desired for the investigation, and what method is compatible with the biological activity of the peptides and proteins.

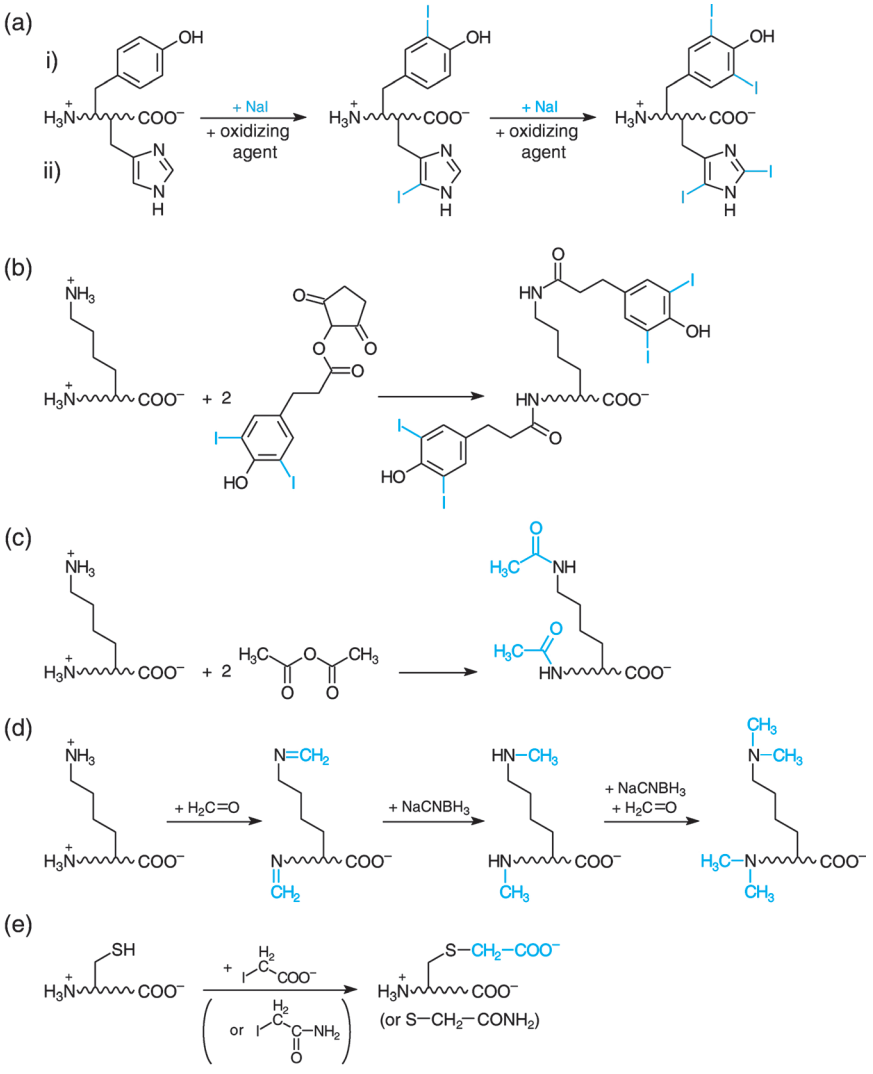


Figure 2.4 Commonly used methods for radioactive labeling of peptides and proteins. (a) Iodination by electrophilic addition: (i) tyrosine, (II) histidine. (b) Iodination of lysine and the N-terminus with the Bolton–Hunter reagent. (c) Acetylation of lysine and the N-terminus by means of anhydride. (d) Alkylation of lysine and the N-terminus using aldehyde and subsequent reduction with sodium cyanoborohydride. (e) Alkylation of cysteine with iodoacetic acid (or iodoacetamide). Source: according to: *Current Protocols of Protein Science*, John Wiley & Sons, Inc., New York, 1995.

The simplest and most common method is radiolabeling with ^{125}I (half-life 60 days) or with ^{131}I (half-life 8 days) in the case that a shorter half-life is desired with a higher specific radioactivity (Ci mmol^{-1}). Iodine is a γ-emitter, which is favorable for detection. Therefore, the strategy for ^{125}iodination of peptides or proteins is discussed below in more detail. By comparison, ^{14}C or ^{3}H labeling have two important advantages: first, different amino acid residues can be modified (lysine or cysteine); second, chemically identical peptides can be prepared by *de novo* synthesis with previously radioactively labeled amino acids. However, the half-lives of ^{14}C and ^{3}H are extremely long (5760 and 12.26 years, respectively) and both are β-emitters.

At this point, the special **safety precautions** and **rules** that are necessary for dealing with radioactively labeled substances and solutions and their strict application is explicitly noted.

2.3.1 Iodinations

The iodination of peptides can be realized directly by the electrophilic addition of tyrosine or imidazole, or indirectly through the insertion of an iodinated tyrosine analogue using the Bolton–Hunter reagent (activated ester) on the free primary amino groups of lysine and the N-terminus (Figure 2.4b). The latter reaction is described in detail in Chapter 6. During the electrophilic addition, $Na^{125}I$ (or $Na^{131}I$) is converted into its activated form ("I^{+}"). This can be realized by oxidizing agents, such as *N*-chlorobenzenesulfonamide (as chloramine-T or immobilized as iodobeads), or more gently for peptides by using lactoperoxidase and H_2O_2. Direct chemical iodination can be carried out twice on the *ortho*-positions of tyrosine and the two imidazole hydrocarbons (Figure 2.4a). Enzymatic iodination is limited to tyrosine. Since lactoperoxidase has a molecular weight of approximately 77.5 kDa, it cannot pass through cell walls or membranes. Therefore, it is used for the careful, selective radioactive labeling of membrane proteins of whole, living cells.

Bolton–Hunter Reagent, Figure 6.2

The most serious drawback of radioactive iodination is that the biological activity is often lost. This is caused either by inactivation of the iodine itself – after all, it occupies the space of a phenyl ring and, thus, can sensitively affect the tertiary structure – or, in certain side reactions, sensitive amino acid residues are oxidized. Tryptophan, methionine, and cysteine are especially oxidation-sensitive amino acids.

Because the iodination is normally incomplete, the radioactively labeled peptides must be separated from the unlabeled starting material. This is done efficiently by reversed phase HPLC (e.g., a RP-C_{18} column) using an acetonitrile gradient for elution. In this way, labeled and non-labeled peptides, as well as different degrees of labeled peptides, can be separated and are available as homogeneous groups.

Reversed Phase HPLC, Section 10.4.2

The reader should be aware that the evaluation of radioactive labeling of proteins and peptides as a quantification method is a special technique for special investigations. A direct comparison with the color and spectroscopic methods described previously is not useful. The latter can be used routinely and without special precautions, they are fast and can be easily used for the quantification of proteins.

Further Reading

Hartree, E.F. (1972) Determination of protein: a modification of the Lowry method that gives a linear photometric response. *Anal. Biochem.*, **48**, 422–427

Noble, J.E. and Bailey, M.J.A. (2009) Quantitation of protein. *Methods Enzymol.*, **463**, 73–95

Sapan, C.V., Lundblad, R.L., and Price, N.C. (1999) Colorimetric protein assay techniques. *Biotechnol. Appl. Biochem.*, **29**, 99–108

Smith, P.K., Krohn, R.I., Hermanson, G.T., Mallia, A.K., Gartner, F.H., Provenzano, M.D., Fujimoto, E.K., Goeke, N.M., Olson, B.J., and Klenk, D.C. (1985) Measurement of protein using bicinchoninic acid. *Anal. Biochem.*, **79**, 76–85

Tsomides, T.J. and Eisen, H.N. (1993) Stoichiometric labeling of peptides by iodination on tyrosyl or histidyl residues. *Anal. Biochem.*, **210**, 129–135

Wiechelman, K.J., Braun, R.D., and Fitzpatrick, J.D. (1988) Investigation of the bicinchoninic acid protein assay: identification of the groups responsible for color formation. *Anal. Biochem.*, **175**, 231–237

Protocols for practical approach

Coligan, J.E., Dunn, B.M., Ploegh, H.L., Speicher, D.V., and Wingfield, P.T. (eds) (2003) *Current Protocols of Protein Science*, John Wiley & Sons, Inc., Hoboken, ch. 3.

Enzyme Activity Testing

3

Hans-Joachim Kytzia

Roche Diagnostics GmbH, DXRHA 6164, Nonnenwald 2, 82337 Penzberg, Germany

3.1 The Driving Force behind Chemical Reactions

Physical-chemical systems are similar to purely mechanical systems, such as Newton's apple falling to the ground, in that they also strive to reach a state of minimum potential energy. This striving to reach a low-energy equilibrium is determined by two different and potentially contradictory tendencies. While potential energy is minimized by virtue of the fact that heat or expansion work is released to the surroundings, entropy – which is a measure of disorder – is maximized.

J.W. Gibbs, in 1878, introduced the state function of "free energy" G to describe this compromise:

$$G = H - TS \text{ in which } H = U + pV \tag{3.1}$$

where H is the enthalpy, which is the sum of the potential energy U and the product of pressure p and volume V reflecting the expansion work. Enthalpy minus the product of entropy S and absolute temperature T yields the free energy.

The change in free energy ΔG is given by the complete integral:

$$\Delta G = \Delta U + p\Delta V + V\Delta p - T\Delta S - S\Delta T \tag{3.2}$$

For the most common condition of a process at constant pressure and constant temperature (i.e., with $\Delta p = 0$ and $\Delta T = 0$) one obtains:

$$\Delta G = \Delta U + p\Delta V - T\Delta S \tag{3.3}$$

The change in free energy is therefore described as the sum of reaction heat ΔU, expansion work $p\Delta V$, and the change in the entropy component $T\Delta S$.

The free energy G and, therefore, the change in free energy ΔG, in terms of their property as functions of state, are dependent only on the state of the particular system and not on the path taken by the system to reach this state. This distinguishes free energy, for example, from work that can be used depending on the path taken or that can simply go unused. A compressed gas that expands can move a piston, for instance, against an external pressure and thereby perform work. If the compressed gas expands into a vacuum, however, no work is performed along the path from the same initial state to the same final state.

The fact that state functions are not dependent on path makes it possible to calculate their change along a hypothetical path from the initial state to the final state without knowing the true path.

All spontaneously occurring processes are characterized by a negative ΔG.

According to the law of conservation of energy, ΔG is equal to zero at dynamic equilibrium, in which the forward reaction and the back reaction are balanced.

Processes with a positive ΔG do not occur spontaneously, although they can be forced by being coupled to further processes having a negative ΔG.

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

Still, the definition of ΔG as a function of potential energy and entropy, both of which are very difficult to determine via experimentation, is not very helpful yet. The value of ΔG is fully realized in its relation – developed by van't Hoff – to the equilibrium constant from the law of mass action:

$$\Delta G = \Delta G^\circ + RT \ln Q \tag{3.4}$$

where R is the gas constant, T is the absolute temperature, and Q is the analog to the mass-action constant with the actual concentrations, not the equilibrium concentrations. The term ΔG° is referred to as standard-state free energy.

At equilibrium, $Q = K$ and $\Delta G = 0$, therefore:

$$-\Delta G^\circ = RT \ln K \tag{3.5}$$

Consequently, it is possible to calculate ΔG° on the basis of the determination of the equilibrium concentrations and to then calculate, on the basis of ΔG°, the actual change in free energy at the actual concentrations. In the case of high concentrations of starting substance and a low product concentration, ΔG can become negative even when ΔG° is positive; it is therefore possible to force a reaction to occur that does not take place spontaneously under standard conditions by selecting the concentrations accordingly.

3.2 Rate of Chemical Reactions

While a negative ΔG is the prerequisite for the spontaneous occurrence of a reaction, conversely, this does not mean that the reaction also actually takes place at a perceptible rate. For instance, a mixture of hydrogen and oxygen is absolutely stable, while the explosion caused by the addition of a warm platinum sheet clearly illustrates that ΔG is negative indeed.

What then determines the rate of a chemical reaction?

Consider a substance A that spontaneously decomposes into B and C; the concentrations of the species are [A], [B], and [C]. The reaction rate v is the rate at which A disappears and/or at which B and C are formed:

$$v = -\frac{\mathrm{d[A]}}{\mathrm{d}t} = \frac{\mathrm{d[B]}}{\mathrm{d}t} = \frac{\mathrm{d[C]}}{\mathrm{d}t} \tag{3.6}$$

In many cases, a simple rate-of-reaction equation is then obtained, according to which the rate is proportional to the concentration of A remaining at the particular point in time:

$$v = -\frac{\mathrm{d[A]}}{\mathrm{d}t} = k[\mathrm{A}] \text{ or } -\frac{\mathrm{d[A]}}{[\mathrm{A}]} = k\mathrm{d}t$$

Integration of this differential equation gives:

$$\ln[\mathrm{A}] = -kt + C \tag{3.7}$$

The constant of integration C results from the basic condition that $[A_0]$ is the initial concentration at time $t = 0$:

$$C = \ln [A_0]$$

therefore:

$$\ln [\mathrm{A}] = -kt + \ln [A_0] \text{ or } [\mathrm{A}] = [A_0]e^{-kt} \tag{3.8}$$

It is therefore possible to determine the constant k, which is the velocity constant, simply by measuring a series of [A]/t pairs. These measurements should yield a constant k. If this is not the case, you could search for a better rate-of-reaction equation that accounts for the influence of further reactants. Often it is possible, however, to vary the reaction conditions in such a way that the simple rate-of-reaction equation can be applied. The rate-of-reaction equation for a monomolecular decomposition would probably be satisfied as well if A would not spontaneously decompose, but rather would hydrolyze in aqueous solution. The concentration of water as the reaction partner, which is very high (55 mol l^{-1}) and is practically unchanged by the reaction, is then incorporated into the velocity constant; the reaction, which is actually complex,

is practically dependent only on the initial concentration of one partner. Since nearly all chemical reactions involve several partners, this principle is applied in most cases.

However, be wary of a common fallacy: The rate-of-reaction equation does not say anything about the mechanism of a reaction and, conversely, the rate-of-reaction equation cannot be derived from the stoichiometry of a reaction.

The velocity constant k is dependent on temperature. Arrhenius postulated a simple law for this, proceeding from the temperature dependence of the velocity constants derived by van't Hoff and the fact that the equilibrium constant is the quotient of the velocity constants for the forward reaction and the back reaction, as follows:

$$\frac{\mathrm{d}\ln k}{\mathrm{d}T} = \frac{E_a}{RT^2}$$

or, assuming that E_a is not temperature-dependent, integrate and find the antilog:

$$k = A\mathrm{e}^{-E_a/RT} \tag{3.9}$$

Therefore, k is dependent on a frequency factor A as a measure of the number of collisions between potential reaction partners and the factor $\mathrm{e}^{-E_a/RT}$, which represents the temperature-dependent portion of the molecules that is capable of entering into the reaction due to its potential energy. Arrhenius called E_a the activation energy. The illustrative significance of this is that a transition state must be passed through on the reaction path, the energy of which is greater by this activation energy than that of the parent compounds and of the reaction products.

3.3 Catalysts

For a chemical reaction to take place at all, ΔG (not ΔG°!) must be negative. For the chemical reaction to proceed at an adequate rate, a sufficient portion of the parent molecules must have enough energy to apply the activation energy.

The effect of a catalyst is that of increasing the rate of a ΔG-favorable reaction by lowering the activation energy; in other words, the catalyst induces the reaction along a different path with a lower activation energy. The value of ΔG is not influenced by this, because ΔG is a path-independent state function. Therefore, the state of the chemical equilibrium is also not changed; the establishment of equilibrium is merely accelerated. Since the rates of the forward reaction and the back reaction are equal in equilibrium, this means that the catalyst accelerates the forward reaction and the back reaction equally.

3.4 Enzymes as Catalysts

In biological systems as well, many reactions take place spontaneously, although at a rate that is much too slow and must therefore be accelerated by means of a catalyst. This is precisely the role that enzymes perform. Enzymes are proteins that typically consist of a few hundred amino acids, only a few of which, however, are directly involved in the bonding and conversion of the substrate molecules. The actual reactive center is often a hydrophobic pocket that encloses the substrate molecule. Amino acid side chains that form hydrogen bridges perform the function of the hydrating water molecules. Due to the shape and size of the pocket, only certain molecules can function as the substrate and, therefore, the shape and size of the pocket determine the typically extraordinarily high specificity of an enzymatic reaction. Although an entire series of amino acids is often involved in the bonding of the substrate, the bonding energy is weak, and is typically much weaker than a single covalent bond. A characteristic feature, however, is a high affinity of the enzyme for the transition state of the reaction to be catalyzed. The mutual interaction between enzyme and substrate has been referred to, since Koshland, as an induced fit: During the substrate bonding, the enzyme molecule takes on a lower-energy conformation, in which the substrate molecule is simultaneously forced into a conformation that approximates the transition state of the reaction to be catalyzed. In sum, these two effects bring about a stabilization of the transition state or, as Arrhenius would have put it, a reduction in activation energy.

In summary, enzymes lower the activation energy of a reaction by stabilizing the transition state. As catalysts, they therefore accelerate the forward reaction and the back reaction without shifting the equilibrium between the substrate and the product.

3.5 Rate of Enzyme-Controlled Reactions

The bonding of the enzyme to a conformation of the substrate that is analogous to the transition state and the resultant reduction in activation energy are obviously coupled to the tertiary and quaternary structure of the enzyme protein. As with all proteins, these are greatly dependent, in turn, on the environment, in particular on the pH and the ionic strength. The dependence of the enzymatic activity as compared to other reactions is striking here, while the temperature dependence of many enzymatic reactions, for instance, is similar to that of normal chemical reactions. To describe an enzymatic activity in a meaningful manner, the experimental conditions must therefore be described in detail.

A further characteristic of the kinetics of enzyme-controlled reactions is the dependence on the substrate concentration. While, in the case of normal chemical reactions, the reaction rate can be increased largely arbitrarily by increasing the concentration of the parent substances, this does not apply to enzymatic reactions. In this case, when the reaction rate is plotted against the substrate concentration, the typical result is a hyperbolic curve that asymptotically approaches a limit value, the so-called maximum rate.

3.6 Michaelis–Menten Theory

In an attempt to explain the hyperbolic dependence of the initial rate of an enzymatic reaction on substrate concentration, Brown and Henry postulated in 1905 that an enzyme–substrate complex is formed in an initial reaction step, which then decomposes in a second step to form the product and the free enzyme. The influence of the back reaction, which is also catalyzed by the enzyme, can be disregarded, since the product concentration is still approximately zero at the beginning of the reaction:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P \tag{3.10}$$

Michaelis and Menten derived a theoretical interpretation of such a process on the basis of two essential assumptions. According thereto, a constant concentration of the enzyme–substrate complex is established in a thermodynamic equilibrium with enzyme and substrate, which is not disrupted by the further reaction to form the product. They further postulated that the reaction of the enzyme–substrate complex to form the product is much slower than the decomposition back to free enzyme and substrate. Secondly, the concentration of free substrate is constant and equal to the total substrate concentration because the substrate concentration is substantially greater than the enzyme concentration.

If [E] stands for the entire enzyme concentration and [S] stands for the total substrate concentration, equaling the free substrate concentration, and [ES] stands for the equilibrium concentration of the enzyme–substrate complex, the free enzyme concentration ([E] – [ES]) can be used to express the dissociation constant for ES (by definition, the reaction of ES to form P does not matter!):

$$\begin{aligned} & k_1([E]-[ES])[S] = k_{-1}[ES] \\ \text{or} \quad & K_S = \frac{k_{-1}}{k_1} = ([E]-[ES])\frac{[S]}{[ES]} \\ \text{solved for [ES]:} \quad & [ES] = \frac{[E][S]}{K_S + [S]} \end{aligned} \tag{3.11}$$

The rate of product formation is:

$$\begin{aligned} & v = k_2[ES] \\ \text{or} \quad & v = \frac{k_2[E][S]}{K_S + [S]} \end{aligned}$$

k_2[E] is obviously equal to the maximum rate V_{max} that is reached when the entire enzyme is present as the enzyme–substrate complex. Therefore:

$$v = V_{max} \frac{1}{\left(1 + \frac{K_S}{[S]}\right)} \tag{3.12}$$

This so-called Michaelis–Menten equation describes the required hyperbolic approach of the reaction rate to the maximum rate as substrate concentration increases: v approaches V_{max} when [S] is a great deal larger than K_S. An obvious significance of the constant K_S also becomes apparent: K_S corresponds to the substrate concentration at which the reaction rate corresponds to half of the maximum rate.

Please note that a common misinterpretation is to regard the constant K_S as a measure of the affinity between enzyme and substrate.

The central assumption of Michaelis–Menten kinetics, namely, that [ES] and [E] are constant, was derived here, based on the assumption that there is a thermodynamic equilibrium between the enzyme, substrate, and the enzyme–substrate complex. This assumption must apply in order for a small K_S to be considered equivalent to a high affinity of the enzyme to its substrate.

The Michaelis–Menten equation can be derived in general, however, for any case of a constant [ES] and [E], for example, brought about according to the steady-state assumption of Briggs and Haldane. Various kinetic causes can underlie the fact that the complex-forming reactions and the complex-consuming reactions remain in equilibrium.

In such a situation, the reaction of the enzyme–substrate complex to form the product does not need to be much slower than its decomposition into enzyme and substrate, and the substrate concentration does not necessarily need to be high relative to the enzyme concentration. In this case, however, the *steady-state* concentration of the enzyme–substrate complex is not equal to the equilibrium concentration and, therefore, is no longer a measure of the affinity between the enzyme and the substrate.

For this typical case, namely, in that it was not demonstrated that [ES] is the equilibrium concentration, one writes K_m (the Michaelis constant) instead of K_S; K_m is simply the substrate concentration that results in the half-maximum reaction rate under the given conditions.

3.7 Determination of K_m and V_{max}

The Michaelis–Menten equation can be applied in various ways to determine K_m and V_{max} from a series of measurements of the initial rate as a function of the substrate concentration.

The direct application of v versus [S] results in the aforementioned hyperbola, for which the asymptotic limit value V_{max} can be calculated using currently available curve-fitting programs. The K_m then results as the substrate concentration associated with $V_{max}/2$ (Figure 3.1).

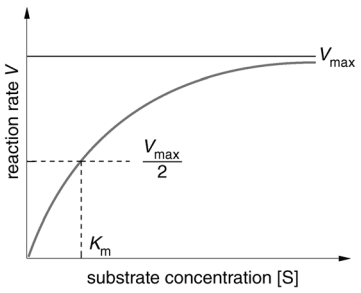


Figure 3.1 Dependence of the reaction rate of an enzymatic reaction on the substrate concentration according to the Michaelis–Menten equation.

Several linear transformations were proposed for the graphic evaluation, of which the double reciprocal plot according to Lineweaver and Burk is perhaps the most obvious and best known, although it may not be best suited for all purposes.

Transformation of the Michaelis–Menten equation yields:

$$\frac{1}{v} = \frac{K_m}{V_{max}} \frac{1}{[S]} + \frac{1}{V_{max}} \tag{3.13}$$

A plot of $1/v$ over $1/[S]$ is therefore a line having a slope of K_m/V_{max} and axis intercepts of $1/V_{max}$ and $-1/K_m$ (Figure 3.2).

The main disadvantage of this plot, the unfavorable distribution of the double reciprocal measurement points in the diagram, and the high weight of the most error-prone measurements associated with the lowest substrate concentrations can be corrected by means of appropriate weighting.

Another method for the graphic determination of K_m and V_{max} was proposed by Eisenthal and Cornish-Bowden (Figure 3.3). They suggested marking the points $-[S]_i$ on the abscissa and the corresponding measurement points v_i on the ordinate. A straight line drawn through the points that belong together results in one line for each pair of measurement points. All of these lines should intersect at the point with the coordinates K_m and V_{max}.

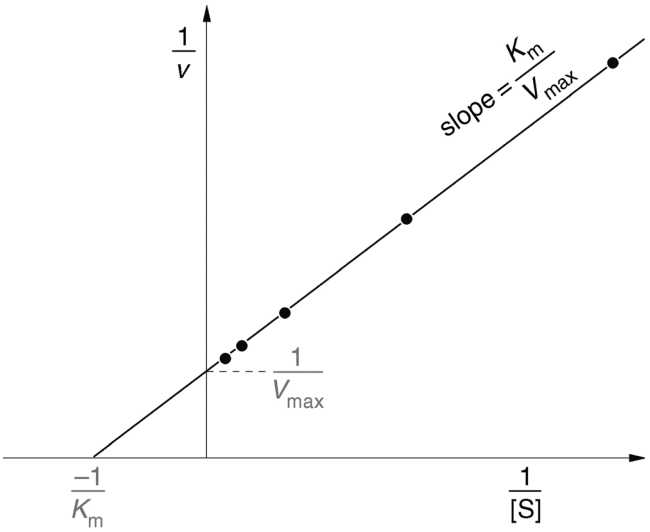


Figure 3.2 Determination of K_m and V_{max} according to Lineweaver and Burk.

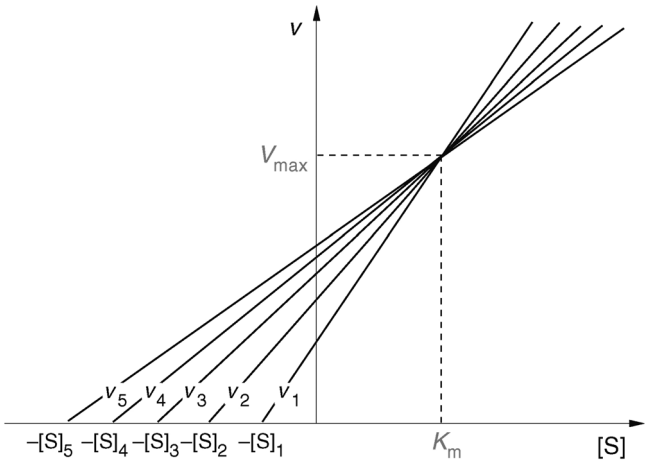


Figure 3.3 Determination of K_m and V_{max} according to the "direct plot" of Cornish-Bowden and Eisenthal. Each line represents one measurement of the reaction rate v_i and the corresponding substrate concentration plotted as $-[S]_i$. The common intersection point of the lines leads to K_m and V_{max}.

3.8 Inhibitors

Inhibitors of enzymatic reactions play such a significant role *in vivo* in the regulation of enzymatic activities and *in vitro* in the determination of metabolic pathways that we will now briefly discuss the enzyme kinetics at least for the main forms of inhibition.

3.8.1 Competitive Inhibitors

Competitive inhibitors compete with the substrate, typically for the same binding site on the enzyme molecule; in this regard they are similar to the substrate or, more effectively, the transition state, but they are not converted into product. Therefore, enzyme can be converted into either the enzyme–substrate complex (ES) or the enzyme–inhibitor complex (EI), but there is no EIS complex. The inhibitor removes enzyme molecules from the equilibrium so that they are no longer available for formation of the ES, but the inhibitor is completely displaced from the enzyme in the case of very high substrate concentration: such an inhibitor therefore does not influence V_{max}. Instead, its effect is to apparently increase the K_m.

In the presence of a competitive inhibitor, the Michaelis–Menten equation becomes:

$$v = V_{max} \frac{1}{1 + \frac{K_m}{[S]}\left(1 + \frac{[I]}{K_i}\right)} \tag{3.14}$$

where [I] is the inhibitor concentration and K_i is inhibitor constant.

In analogy to the definition of the Michaelis constant, the inhibitor constant K_i corresponds to the inhibitor concentration that reduces the reaction rate to 50% at the particular substrate concentration.

Given a constant inhibitor concentration and an increasing substrate concentration, v approaches V_{max}; given a constant substrate concentration and an increasing inhibitor concentration, v approaches zero.

If the dependence of the reaction rate on the substrate concentration is determined in a series of experiments using different inhibitor concentrations, the Lineweaver–Burk plots yield a series of lines, all of which intersect at point $1/V_{max}$ on the ordinate. The slope of the lines is then:

$$\frac{K_m\left(1+\frac{[I]}{K_i}\right)}{V_{max}}$$

which becomes:

$$\frac{K_m}{V_{max}} \quad \text{for } [I]=0$$

3.8.2 Non-competitive Inhibitors

A non-competitive inhibitor binds to the enzyme without interfering with the binding of the substrate. In addition to the enzyme–substrate complex ES, the enzyme–inhibitor–substrate complex EIS is then also present, which decomposes to product at a slower rate or, in the extreme case, not at all. Such an inhibitor does not influence K_m. Instead, it reduces V_{max}.

The derivation of the Michaelis–Menten equation for this case yields:

$$v=\frac{V_{max}}{1+\frac{[I]}{K_i}}\times\frac{1}{1+\frac{K_m}{[S]}} \tag{3.15}$$

As the inhibitor concentration increases, V_{max} therefore is reduced to the rate of the breakdown of EIS; for $i=0$, the result is the unmodified Michaelis–Menten equation.

The inhibitor constant K_i can again be determined from the dependence of the reaction rate on the substrate concentration at various inhibitor concentrations. In the Lineweaver–Burk plot, a family of lines results that intersect at the point $-1/K_m$ on the abscissa. K_i is determined from the ordinate intersection $(1 + i/K_i)/V_{max}$ or from the slope:

$$\frac{K_m\left(1+\frac{[I]}{K_i}\right)}{V_{max}}$$

In addition to these two pure main forms, there are also several other inhibition mechanisms. For example, a non-competitive inhibitor can influence V_{max} as well as K_m. The double reciprocal plot is very well suited – provided the measurements are sufficiently precise – for detecting such forms on the basis of the position of the point of intersection of the family of lines.

3.9 Test System Set-up

The measurement of enzymatic activities can be appropriate for various purposes, such as for investigating the physiological function of the enzyme, its embedding in metabolic pathways, or its role in metabolic disorders. The high amplification factor makes it possible, by measuring representative enzymes, to detect traces of contamination or, in clinical chemistry diagnostics, to detect the release of enzymes from damaged organs long before other symptoms occur. The enzymatic reaction can also be used, of course, to determine the concentration of a substrate. The extraordinary specificity of many enzymatic reactions makes it possible, for instance, to perform a precise and interference-free determination of the concentration of ammonia in blood, which is in the micromolar range. The difficulties encountered in measuring enzymatic reactions are often associated with the low stability of the sensitive protein molecules and

with the high dependence of the enzymatic activity on the external conditions. These dependencies must be known and the critical influences must be controlled, therefore, in order to design a robust and simple system for measuring a certain enzymatic activity. The systematic optimization and fine-tuning of an enzyme test preferably starts with the definition of the particularly critical conditions, for example, in the order described in the following.

3.9.1 Analysis of the Physiological Function

Enzymes have been optimized, through millions of years of natural selection, to perform their function as well as possible in a very specific environment. When these conditions are known, essential enzymatic properties can be predicted. For example, when it is known that an enzyme is of lysosomal origin it can be assumed that the optimum pH is close to pH 4 and that the optimal enzymatic activity could be at a relatively low ionic strength. An enzyme that accepts the *N*-acetylglucosamine group as part of the substrate, but does not accept glucose derivatives instead, is very likely to be competitively inhibited by acetate buffer, since it is most likely to contain a binding site for this residue. In addition, the known narrow regulation of glutamate dehydrogenase (EC 1.4.1.3) in mammals by means of the ADP-to-ATP ratio, among other things, as a measure of the energy charge in the cell can be predicted on the basis of its position at the interface of essential metabolic pathways, namely, as a supplier of intermediates for amino acid metabolism, energy metabolism in the citric acid cycle, and for ammonia excretion in the urea cycle. The same is true for the fact that both NADH and NADPH are accepted as the second substrate. By contrast, anabolic and catabolic functions are performed by different enzymes in many bacteria. This is a clear indicator that the regulation of the purely anabolically active GlDH (EC 1.4.1.4) is much less complex indeed, and also that this enzyme can only convert NADPH, but not NADH.

Clearly, a careful analysis of the information available in the literature can give valuable hints and may help in making proper choices instead of finding oneself caught in dead-end streets.

3.9.2 Selecting the Substrates

The main criterion for selecting a suitable substrate is the specificity of the enzyme to be investigated. The natural substrate should be used whenever possible. If there is no choice but to use an artificial substrate, the risk of detecting a secondary activity of the enzyme to be investigated or even a contaminant in the enzyme preparation must be ruled out.

Further criteria are, of course, the availability in a characterized form, the stability under incubation conditions, and suitability for the intended detection system.

It is also worth asking if the disappearance of the parent substance (decreasing reaction) or the formation of the product (increasing reaction) should be measured. On the one hand, increasing reactions are less problematic to monitor; on the other hand, even in the case of highly complex enzymatic mechanisms, the parent substance will disappear immediately, while the appearance of the product may be initially delayed.

3.9.3 Detection System

Reaction rates can be measured in two fundamentally different ways. Continuous measurements track a property without intervening in the reaction system, while discontinuous measurements involve taking samples at certain intervals and analyzing them. To do this, the reaction in the sample that was taken must be abruptly stopped. Continuous measurement provides substantial advantages, in particular in measuring the time dependence of the reaction rate and in the extrapolation to the initial rate, that is, to a point in time at which the composition of the reaction mixture is exactly known and for which interferences caused by substrate shortage, product inhibition, or the like are ruled out.

On the other hand, discontinuous measurement is the preferred option in cases where enzymatic activity and assessment of the indicator exclude each other. This is true for many cases of enzyme reactions taking place at neutral or slightly acidic pH that yield a chromogenic or fluorogenic indicator that becomes “visible” only in alkaline pH.

3.9.4 Time Dependence

For many enzymatic reactions, the reaction rate decreases continuously as the reaction time progresses. This may be due to the progression of the reaction itself, such as a substrate shortage (the concentration of ES is no longer constant), or to the product concentration, which increases over time and results in a disruption due to the back reaction or product inhibition. External causes such as inactivation of the enzyme by temperature, pH, or the like are also possible. In this case, a sufficiently long window of time needs to be identified in which the reaction rate is independent of time. Alternatively, it may be worth performing an extrapolation to the initial reaction rate at time zero.

In contrast to this behavior, one often observes a sigmoidal interrelationship between the product formation and time: the reaction rate represented by the slope of the plot is initially close to zero, then reaches a maximum, and, finally, decreases. Such an interrelationship is a clear sign that the simple models do not correctly depict the true course of the enzymatic reaction. Many different causes can be involved. It is possible that, at the beginning, the rate of product formation is not equal to the rate of substrate consumption, because the reaction takes place via intermediate steps; it is possible that the active form of the enzyme is formed only by means of a conformation change under the test conditions or by combination with a coenzyme. In each case, suitable means must be implemented to prevent such behavior. It is often sufficient to equilibrate the enzyme under test conditions and then add only the substrate in order to start the reaction.

3.9.5 pH Value

In contrast to simple chemical reactions, the pH value determines the rate of the enzymatic reactions not only in that H^+ or OH^- ions enter into the rate-of-reaction equation as reaction partners. In addition, the overall charge of the protein and that of individual amino acids reversibly determine the conformation of the enzyme. Conformational changes induced by pH-shifts are generally reversible within wide limits, but they have a strong effect on the binding and conversion of the substrate. Enzymatic reactions are therefore characterized by an optimum pH; the reaction rate decreases on both sides of the optimum.

An irreversible inactivation of the enzyme at certain pH values is also possible. This can become a problem primarily when the enzyme functions far outside of its optimum pH, for example, within a reaction cascade. Incubation at a pH close to the isoelectric point, where the enzyme no longer carries a net electric charge, particularly favors denaturation and precipitation of the protein. An irreversible inactivation is usually detected when the enzyme is preincubated at a critical pH and then, after rebuffering, the original activity is no longer reached even at the optimum pH. But care should be taken before such an inactivation is considered to be irreversible: reconstitution after a reversible conformational change can take a great deal of time (from seconds to hours!).

It can make sense to initially roughly estimate the optimum pH of the enzyme in steps of approximately 0.5 pH units and to then determine the optimum pH more accurately in an experiment having smaller increments. When performing the rough estimate, the pH curve often must be composed of segments determined with various buffer substances. These segments should overlap one another so that the possible effect of simply changing the buffer system can be corrected.

Possibly, however, it is not only the rate of the enzymatic reaction that is pH-dependent but the signal of the indicator system as well, and so the result must be corrected accordingly. For example, the absorption of *para*-nitrophenol doubles in the pH range between pH 7 and 9.

3.9.6 Selecting the Buffer Substance and the Ionic Strength

The buffer is selected so as to hold the pH value of the reaction mixture constant at the optimum pH. To this end, the buffer capacity of the sample and pH shifts caused by the progression of the reaction must be compensated for. In contrast, the ionic strength in the formulation often must be held low, and therefore it is worth choosing the buffer carefully.

The capacity of a buffer at a given pH is calculated using the pK_a value, which is found in tables, according to the Henderson–Hasselbalch equation:

$$pH = pK_a + \log\left(\frac{[A^-]}{[HA]}\right) \tag{3.16}$$

where the sum of $[A^-]$ and [HA] is, of course, equal to the total concentration of buffer.

As a rule of thumb, pK_a and the pH value should not differ by more than 1.

Furthermore, the selected buffer should be inert to the enzyme, active components, and the detection system in the applicable system. It should not form precipitates with test or sample components nor chelates with essential test components such Ca^{2+} or Mg^{2+}, and it should have the lowest possible ionic strength.

Good's buffers are selected especially for use in biochemical systems. First described in 1966, the selection of Good's buffers has expanded considerably by additional similar compounds (including TRIS, for example, which is not one of the original Good's buffers but has become the probably most common) and now provides a solution for nearly every pH range. The main disadvantage of these substances is the relatively high temperature dependence of the pH value of up to 0.03 units per degree. In addition, the original Good's buffers had been selected for not being membrane-permeant, but a few of the derivatives are and can therefore cause the pH gradients in compartments to break down.

Similarly to the pH value, the ionic strength also influences the enzymatic activity, although usually to a much smaller extent, and often negatively. It is advantageous to select the buffer concentration to be as low as possible and to adjust the optimum overall ionic strength using salt.

3.9.7 Temperature

Many enzymatic reactions follow a rule of thumb derived from the Arrhenius' equation that predicts a doubling of the reaction rate when the temperature is increased by 10 °C, although this may be surprising due to the relative complexity of the enzymatic reactions. The reaction temperature must therefore be well controlled and held constant. The limitations of the equipment that is available will substantially influence the choice of the measurement temperature particularly in the case of continuous measurement in an appropriate measuring device.

The temperature can be increased in order to increase the turnover. This does not mean that the critical signal-to-noise ratio will be improved, however, since unspecific secondary reactions are usually accelerated in the same way.

Beyond a certain limit, the reaction rate will start to drop again when the temperature is further increased: The enzyme is inactivated, as in the case of an extreme pH value, and usually irreversibly so. The optimum temperature is highly dependent on the chemical environment, particularly on pH and on the ionic strength. Therefore, the incubation temperature should not be specified until the remaining conditions have been defined.

3.9.8 Substrate Concentration

Both K_m and V_{max} can be determined only after all the reaction conditions have been defined exactly; both constants depend on all the aforementioned variables.

Once K_m is known, the undefined reaction and measurement conditions can be finalized, primarily the substrate concentration. A commonly stated rule of thumb is that the reaction velocity should be 90% or more of the maximum rate, which can be accomplished by selecting a substrate concentration of more than $10K_m$. The extent to which this is really necessary must be decided in the specific case in question. In any case, it is important, however, to select the substrate concentration to be so high that the substrate shortage that sets in as the reaction progresses results in an insubstantial change in the reaction rate over the observation period. A suitable countermeasure is to shorten the measurement window, provided this is possible in light of the sensitivity of the determination.

3.9.9 Controls

In most cases, further unspecific effects will superpose on the actual enzymatic reaction to be measured, and these must be eliminated by means of suitable control measurements. For example, the reaction will also take place in the absence of the catalyzing enzyme (e.g., spontaneous hydrolysis) at a rate that is very slow but is usually not negligible. The reaction rate must then be subtracted, as the reagent blank, from the signal that is measured, under the exact same conditions although in the absence of the enzyme. In addition, biological samples to be investigated for enzymatic activity often contain components that, under assay conditions, give a response that is indistinguishable from that of the enzymatic reaction. In such cases, it can be necessary to take a sample blank as the second control by performing the measurement with the sample but in the absence of another key component (e.g., the substrate). If it is not possible to keep the unwanted secondary reactions within a negligible order of magnitude – for example, by using inhibitors, additional helper enzymes, or trapping reactions – then it must be ensured by means of control measurements that every interfering influence is detected and is corrected exactly once. Designing a suitable system of controls can then become a secondary task that is just as comprehensive and complex as the optimization of the actual test.

Further Reading

Bergmeyer, H.U., Grassl, M., and Bergmeyer, J. (eds) (1983) *Methods of Enzymatic Analysis*, 3rd edn, vol. I, VCH-Verlagsgesellschaft, Weinheim.

Copeland, R.A. (2000) *Enzymes: A Practical Introduction to Structure, Mechanism, and Data Analysis*, 2nd edn, Wiley-VCH Verlag GmbH, Weinheim.

Dixon, M. and Webb, E.C. (1979) *Enzymes*, 3rd edn, Longman, London.

Fersht, A. (1998) *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, W.H. Freeman, New York.

Microcalorimetry

4

Alfred Blume

Martin-Luther-University Halle-Wittenberg, Faculty of Chemistry, Physics and Mathematics, Institute of Chemistry - Physical Chemistry, von Danckelmann Platz 4, 06120 Halle/Saale, Germany

Calorimetry is a method used to determine the exchange of heat between a closed system and its surroundings. The method is already quite old and was based on initial observations at the end of the eighteenth century that ice takes up heat during melting without changing its temperature. The term "latent heat" was coined for this effect (Joseph Black, Edinburgh). Most of the calorimetric measurements performed at that time had the goal of understanding chemical and physical processes. However, the first applications to biological systems were also performed. Lavoisier, for instance, built around 1780 the first ice calorimeter devised for determining the metabolism of a guinea pig by measuring the amount of molten ice. This apparatus was based on the insight of Joseph Black about the latent heat (heat of melting) of ice.

Calorimetry became soon a widespread method in chemistry and physics for the determination of thermodynamic properties of materials. Comprehensive data on specific heats, heats of reaction, and heats of phase transitions were collected and tabulated. Calorimetry as a useful tool to study biological systems was only recognized much later and more extensive applications appeared only in the last 20–30 years. The technical development of the mechanical and electronic hardware was a prerequisite for this progress, because higher sensitivity is needed for the thermodynamic characterization of biological samples due to the fact that sometimes only low amounts of material are available. Because of this development of more and more sensitive instruments needing minute amounts of sample the term "microcalorimetry" was introduced.

In calorimetry the heat arising during a reaction and exchanged between the closed system and its surroundings is measured under quasi-isothermal conditions. When heat is evolved in the sample, the reaction is called exothermic and the heat flows to the surroundings. In an endothermic reaction the heat flows from the surrounding into the closed system. Several methods exist to determine the amount of heat exchanged between the system and its surrounding which will not be described here.

Apart from determining the heat of reaction inside a system under quasi-isothermal conditions (isothermal calorimetry) another type of calorimetric experiment can be performed where physical processes are triggered by changes in temperature. Prime examples for these processes are the already mentioned melting of ice or the evaporation of fluid water at the boiling point. For the melting or the evaporation process, energy is needed that is supplied by the flow of heat from the surroundings. In biological systems, processes induced by changes in temperature are, for instance, the unfolding and denaturation of proteins in aqueous solution, the "melting" of the DNA double helix or the transition of membranes from an ordered into a disordered state. This type of calorimetry is called scanning calorimetry because the sample is usually heated and cooled between high and low temperature.

The applications of calorimetric methods to biological systems such as proteins, nucleic acids, or membranes in aqueous buffer systems is closely related to the development of highly sensitive differential scanning calorimeters (DSCs) in which a sample and a reference are

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

scanned in temperature at the same time. Two scientists were at the forefront of the developments of these instruments, namely, Julian Sturtevant and Peter Privalov. Their contributions led to the first commercially available DSC instruments, first based on the design of P. Privalov and manufactured in Russia. Later on, DSC instruments were also developed by J.F. Brandts and sold by the company MicroCal, USA.

Based on a revolutionary idea by J.F. Brandts, the MicroCal DSC instruments were later converted into isothermal titration calorimeters (ITCs). This was achieved by inserting into the sample cell a motor driven syringe where the injection needle had a stirring paddle. The solution was stirred by rotation of the syringe and the reactant was injected by driving the plunger with a step motor, thus injecting between 5 and 25 µl step by step. The instrument operates in the power compensation mode under quasi-isothermal conditions. This led to a drastic increase in sensitivity and particularly to strongly reduced measuring times compared to the previously used heat flow calorimeters. With this instrument it was possible to perform binding studies of ligands to proteins with high precision. As mentioned above the method is called isothermal titration calorimetry though in reality the sample and the reference cell are constantly heated with low power to enable a power compensation mode for exothermic as well as endothermic processes. The change in temperature during an ITC experiment lasting between two and three hours is, however, negligible. All commercially available DSC and ITC instruments for biological systems are based on the same principle and differ only in the electronics, software, and the design of the injection syringes. Automated systems using robotics for cleaning and supplying multiple samples are also sold.

In the following sections the basics of calorimetry will be outlined and examples of the application of the two methods DSC and ITC to biological systems and the analysis of the data will be presented.

4.1 Differential Scanning Calorimetry (DSC)

When the sample cell of a differential scanning calorimeter has been filled with an aqueous solution of, for instance, proteins, nucleic acids, or suspensions of lipid vesicles, and the reference cell has been filled with buffer, both cells are heated with a constant power to achieve an increase in temperature. This temperature increase triggers endothermic transitions in the dissolved molecules, which are recorded by the instrument. In protein solutions a change in secondary structure from the native to the denatured state is induced. This unfolding can either occur as a two-state process or via different intermediate conformational states. For double stranded DNAs the temperature increase induces the double-helix to coil transition. This transition can also occur with different sequential intermediate states or in an all-or-none process.

In the case of lipid vesicles or liposomes as models for biological membranes a temperature increase leads in many cases to a change in the order and conformational states of the lipids in the bilayer membrane. At low temperature, in the so-called "gel phase", the lipids are highly ordered and the alkyl chains are in an all-*trans* conformation. An increase in temperature leads to a "melting" of the chains, the bilayers convert into the liquid-crystalline state, in which the order of the chains is reduced by the appearance of *gauche*-conformers in the chains. Concomitantly, the mobility of the molecules is drastically increased so that lateral diffusion becomes very fast. The temperature and the temperature range at which this process occurs depend on the chemical structure of the lipids and the composition of the lipid mixture. Lipids with highly unsaturated or branched chains are usually already in the liquid-crystalline state at room temperature or a temperature above the freezing of water, so that no endothermic transition is observable by increasing the temperature. When phospholipids are mixed with cholesterol, a molecule that occurs in high percentages in biological membranes, complex changes in the thermotropic behavior are observed. Generally, the endothermic transition is broadened and disappears completely at high percentages of cholesterol. The lipids are transformed into a so-called "liquid-ordered" state, in which the chains are more in an all-*trans* conformation, but the rotational and lateral diffusion remains fast.

For investigations of thermally induced transitions a DSC instrument must fulfil the following requirements:

1. The instrument has to have a high sensitivity, so that transitions in diluted aqueous solutions can be recorded with sufficient precision. Particularly for protein solutions this is essential as

higher protein concentrations in an aqueous solution can lead to protein aggregation, which has to be avoided.

2. The cell volume must be as small as possible to reduce the amount of material needed.
3. The reproducibility of the baseline, recorded with both cells filled with water or buffer, has to be given. If this is not the case, the integration of transition peaks extending over a wide temperature range becomes unreliable. Usually, this baseline is subtracted from the experimental curve before integration.
4. Heating and cooling rates should be precisely controllable. The sensitivity in the cooling mode should be similar to that in the heating mode. This is particularly important for reversible transitions when, for instance, the refolding enthalpy of a protein needs to be determined.

These requirements are usually only fulfilled in instruments working in the adiabatic or quasi-adiabatic mode and using the differential principle with two cells with power compensation. In addition, these instruments have fixed and non-removable cells, which are filled from the outside. This type of design and construction is essential for achieving highly reproducible baselines. Figure 4.1 shows the principle construction of a DSC instrument with power compensation.

A typical DSC instrument has two cells, a reference and a sample cell. Both are enclosed in an adiabatic shield to prevent heat flux from and to the cells from the surroundings. Both cells and the adiabatic shield are heated in a controlled way, yielding the desired heating rate by heaters attached to the cells and the adiabatic shield. The temperature of the cells and the adiabatic shield are measured by temperature sensors. These are thermoelements or semiconductor thermosensors that detect temperature differences between both cells and the surrounding adiabatic shield. The signals from the thermosensors are used to adjust the heating power so that the temperature difference is reduced to almost zero. For an endothermic transition in the sample cell, part of the supplied heat is used for the endothermic transition so that the temperature of the sample cell lags behind. This leads to a temperature difference between sample and reference cell that is detected and the heating power is increased by the controller. The additional heating power P_{diff} needed to maintain a zero temperature difference is stored in the computer as a function of time and/or temperature and also displayed on the computer screen. This curve is the actual thermogram, which is later evaluated by integration of the endothermic peak. In an ideal DSC instrument both cells filled with a solution or buffer have the same heat capacity so that P_{diff} should be zero when the cells are heated to record the baseline. However, this is normally not the case as due to difficulties in the manufacturing process the cells usually have slightly different heat capacities, which also change with temperature. Therefore, the baseline is rarely a horizontal straight line at $P_{diff} = 0$ but a curve characteristic

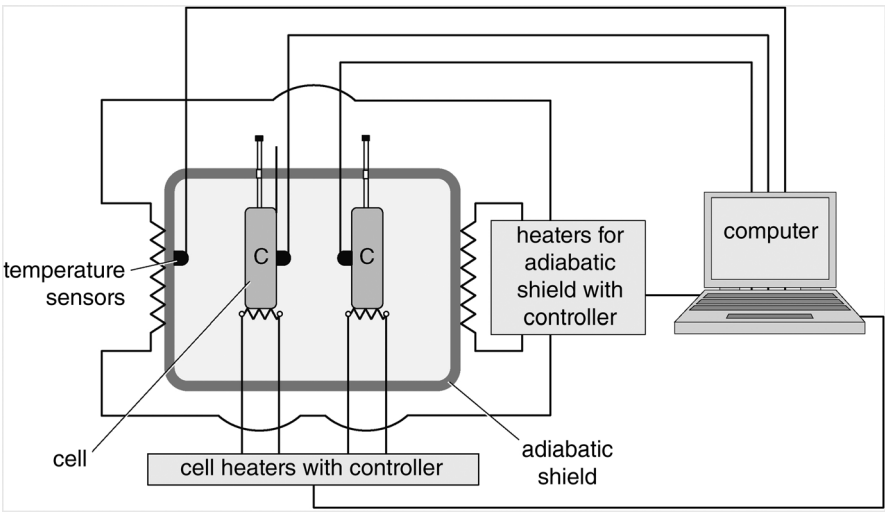


Figure 4.1 Scheme for the design of a differential scanning calorimeter. The two identical cells (C) are heated with the same constant power. Temperature sensors detect any temperature difference occurring between the cells. The signal is fed to a controller which adjusts the heating power of either one of the cells to reduce the temperature difference to zero. Both cells are surrounded by an adiabatic shield that is also heated. Its temperature is measured by temperature sensors and a controller adjusts the heating power so that the shield temperature is identical to the cell temperature to avoid heat flow from or to the cells from the surrounding shield. All heating powers are computer controlled. For an endothermic transition the additional power needed to maintain the sample cell at an identical temperature to the reference cell is recorded and displayed on the screen as a function of cell temperature.

for each instrument. What is important is that this baseline is reproducible and stable over time. To account for changes the baseline is normally recorded at regular intervals and stored in the computer before making the actual experiments. Before evaluation of the thermograms the stored baseline is subtracted from the new experimental curve. Commercial DSC instruments are equipped with two different types of cells, the so-called lollipop cells and the capillary cells. The lollipop-type cells, which are slightly thicker than a coin, have a capillary for filling the cell attached to them. This capillary is fed through the outer adiabatic shield so that the cell can be filled from the outside. Cells of the capillary type are constructed from several loops of a metal capillary fed through the adiabatic shield and open at both ends, so that the cell can be filled in the flow-through mode from the outside. The cell volume of commercial instruments varies between 0.1 and ca 1 ml. The material of the cells is usually gold, platinum, or tantallum alloy. Instrument versions with automatic filling and cleaning accessories also exist. These versions can be connected to a robotic system with a 96-well plate to increase the throughput of the system in terms of the systematic variation of the cell content.

The commercially available DSC instruments have nowadays excellent sensitivity and baseline stability. The short time noise of the baseline is 0.015–0.025 μW. This is equivalent to 0.21–0.35 μcal s^{-1} using a heating rate of 1 °C min^{-1}. Also very important is the repeatability of the baseline after refilling the cell. Here the deviations are characteristically 0.025–0.5 μW, with the instruments using capillary cells having lower deviations than those using the lollipop shaped cells.

As already mentioned above the differential heating power is recorded as a function of time or temperature when the DSC is operated using a constant heating or cooling rate. This differential power P_{diff} depends on the difference in heat capacity c_{diff} between the two cells:

$$c_{diff} = c_{probe} - c_{reference} = P_{diff}\frac{dt}{dT} \tag{4.1}$$

where dt is the time interval for a change in temperature dT.

After recording the baseline when both cells are filled with buffer, the sample cell is filled with the solution of the protein or a lipid suspension, for instance, and a second curve is recorded. Figure 4.2 shows schematically the two curves obtained for this case.

Figure 4.2 shows a buffer–buffer baseline and the thermogram of a suspension of hydrogenated soy bean phosphatidylcholine vesicles that undergo a transition from the ordered lamellar gel phase to the liquid-crystalline phase. The shift of the experimental curve relative to the baseline is clearly evident. This is due to the different heat capacity of the sample cell when it is filled with the vesicle suspension. From this shift the apparent molar heat capacity ${}^{\phi}C_p$ of the suspended lipid can be determined provided the specific volume of the lipid V_L is known. The relevant equation connecting these quantities with the experimentally observed shift in the baseline is shown in Equation 4.2, where V_W is the specific volume and c_{pW} the specific heat of water and m_L and M_L are the mass of lipid in the cell and the molar mass of the lipid,

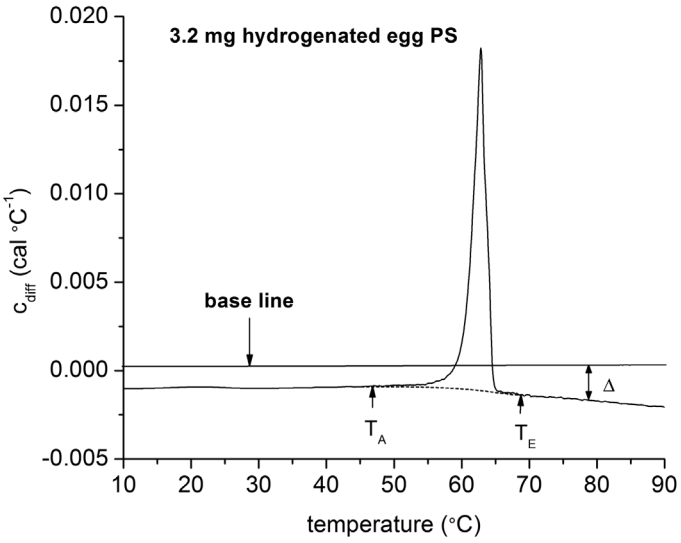


Figure 4.2 DSC curve of a suspension of liposomes of hydrogenated egg phosphatidylserine (PS) vesicles, where a phase transition of the bilayer membrane between the so-called gel phase and the liquid-crystalline phase is occurring. The second curve shown is the baseline when both cells are filled with buffer or water.

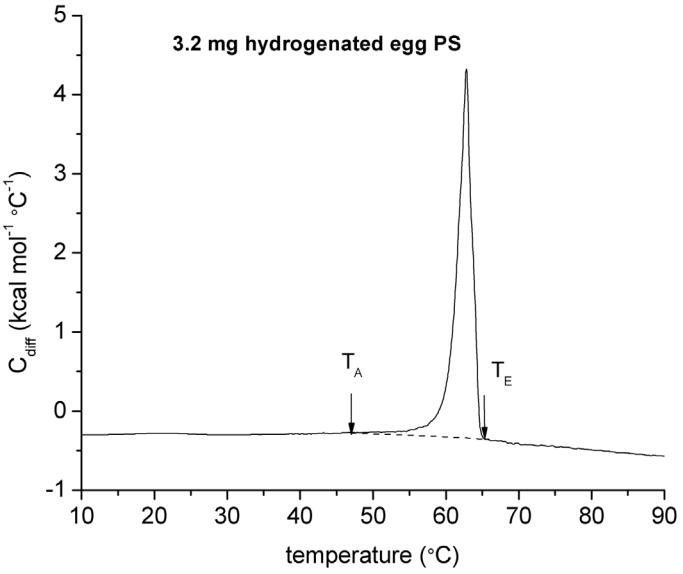


Figure 4.3 Normalized DSC curve after subtracting the baseline and normalizing to the number of moles of substance in the cell. T_A and T_E are the onset and end temperatures of the transition, respectively, that is, the temperature range for the integration of the peak used to determine the molar transition enthalpy ΔH_{trans}.

respectively:

$$^{\phi}C_p = \left[c_{pW}(V_L/V_W) - \Delta/m_L\right]M_L \tag{4.2}$$

The equation is also applicable on exchanging the relevant terms for protein or nucleic acid solutions. Thus the apparent molar heat capacity $^{\phi}C_p$ can be determined, which is an important thermodynamic quantity. If only the baseline as shown in Figure 4.2. is subtracted and the resulting curve is normalized by considering the concentration in the cell and its volume the normalized thermogram C_{diff} as a function of temperature is obtained as shown in Figure 4.3.

Figure 4.3 also shows the two temperatures T_A and T_E which are, respectively, the onset and end temperatures used for the integration of the peak to determine the molar transition enthalpy. The integration has to be performed using the interpolated dashed baseline C_0, which has to be set or estimated by the experimenter. The molar transition enthalpy then corresponds to the area between the C_{diff} curve and the baseline C_0 in the transition region:

$$\Delta H_{conversion} = \int_{T_A}^{T_E} (C_{diff} - C_0)\,dT \tag{4.3}$$

A completely analogous evaluation of the thermograms is possible for solutions of proteins or nucleic acids where temperature induced denaturation or conversion from double-helix into coil are observed, respectively. Figure 4.4 shows the denaturation of lysozyme occurring at a defined temperature T_m. In the case of protein denaturation a commonly observed phenomenon

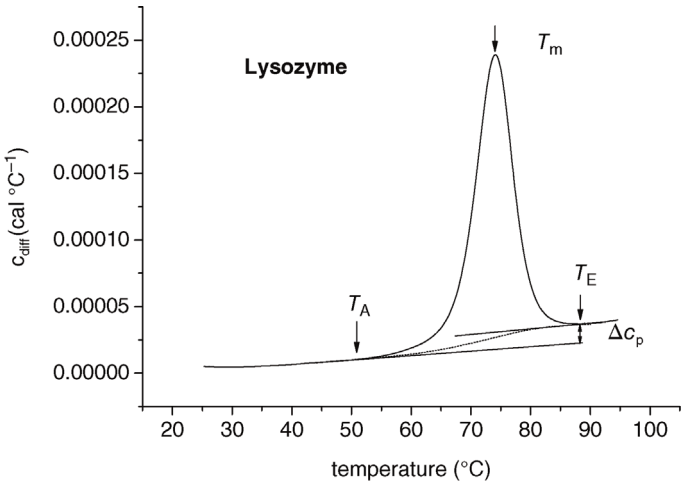


Figure 4.4 DSC curve of a solution of lysozyme (not normalized). The thermally induced denaturation occurs at a defined temperature (T_m) and can be described using a two-state model.

is a shift in the baseline to higher values of C_{diff} after the transition peak. This complicates the integration of the transition peak. It is common practice to construct a sigmoidal interpolated baseline between T_A and T_E as shown in Figure 4.4. The sigmoidal shape is derived from a first integration and is therefore proportional to the integral over the transition peak.

The two-state model is a simple and accepted model for the denaturation of a protein and for a transition between two different lamellar phases of a lipid. For the unfolding of a protein in solution, an equilibrium is assumed between the two states, namely, the native and the denatured state:

$$A \rightleftharpoons B \tag{4.4}$$

Using the law of mass action, the equilibrium constant K is defined as the ratio of the concentrations of the two species:

$$K = \frac{[B]}{[A]} = \frac{\theta}{(1-\theta)} \tag{4.5}$$

where θ is the degree of transition between the two states, running from 0 to 1. The relation between θ and the equilibrium constant K is thus:

$$\theta = \frac{K}{1+K} \tag{4.6}$$

The equilibrium constant K depends on temperature. Its temperature dependence is described by the van't Hoff equation containing the van't Hoff transition enthalpy $\Delta H_{v.H.}$. One obtains for $K(T)$ therefore the following equation:

$$K(T) = K(T_m)\exp\left[\frac{-\Delta H_{v.H.}}{R}\left(\frac{1}{T}-\frac{1}{T_m}\right)\right] \tag{4.7}$$

Inserting Equation 4.7 into Equation 4.6 leads to an equation describing the degree of transition θ as a function of temperature. From the slope of the θ versus T curve at the midpoint of the transition, at the temperature T_m, the van't Hoff transition enthalpy $\Delta H_{v.H.}$ can be obtained:

$$\left(\frac{d\theta}{dT}\right)_{T_m} = \frac{\Delta H_{v.H.}}{4RT_m^2}; \quad \Delta H_{v.H.} = 4RT_m^2\left(\frac{d\theta}{dT}\right) \tag{4.8}$$

Differentiation of Equation 4.6 and multiplication with the calorimetrically determined transition enthalpy ΔH_{cal} leads to Equation 4.9, which describes the theoretical curve for c_{diff} for a two-state transition:

$$c_{diff} = \Delta H_{cal}\left(\frac{d\theta}{dT}\right) = \Delta H_{cal}\left(\left\{1+\exp\left[\frac{-\Delta H_{v.H.}}{R}\left(\frac{1}{T}-\frac{1}{T_m}\right)\right]\right\}^{-2} \times \exp\left[\frac{-\Delta H_{v.H.}}{R}\left(\frac{1}{T}-\frac{1}{T_m}\right)\right]\frac{\Delta H_{v.H.}}{RT^2}\right) \tag{4.9}$$

Figure 4.5 shows schematically the different calculated curves obtained when the ratio $\Delta H_{v.H.}/\Delta H_{cal} = n$ is varied. For $n = 1$ the lowest curve is obtained. This curve applies to an unfolding

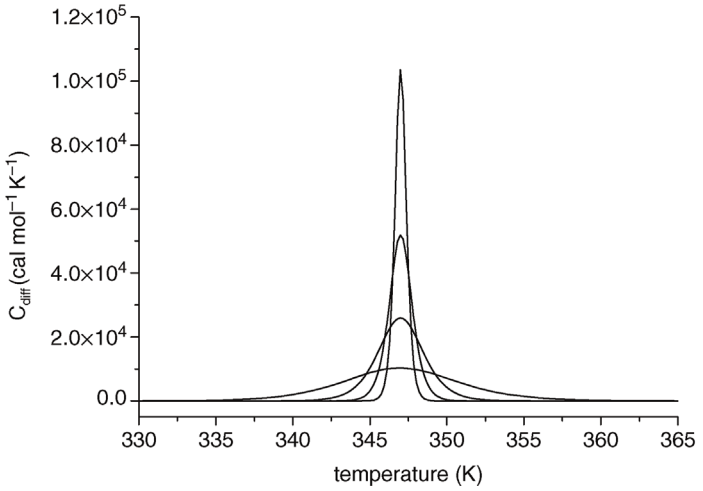


Figure 4.5 Calculated curves using the two-state model with different values for $n = \Delta H_{v.H'}/\Delta H_{cal}$. The value of n was varied between 1, 2.5, 5, and 10. The value of ΔH_{cal} was 100 kcal mol^{-1} in each case. Increasing n leads to a narrowing of the transition, while the area under the peak stays constant.

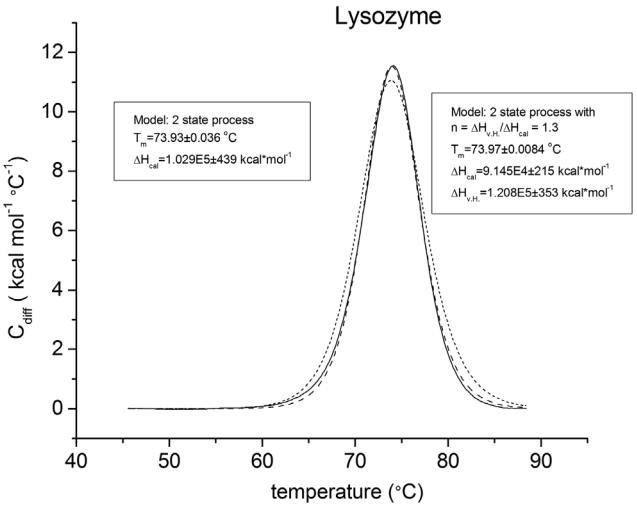


Figure 4.6 Normalized DSC curve (black) for the denaturation of lysozyme together with two different simulated curves based on the two-state model with $n = 1$ (short dashed curve) and $n = 1.3$ (dashed gray curve).

transition for a protein from a native to the denatured state, which is frequently observed. If the ratio n is larger than 1, the transition is of a cooperative nature. This is the case for lipid bilayer transitions where many molecules cooperatively go from the gel to the liquid-crystalline state upon heating. The quantity n is also called the cooperative unit. For transitions of small lipid vesicles the cooperative unit n is between 50 and 100. For multilamellar vesicles with negligible curvature n becomes larger, because the bilayers have less curvature, which increases the cooperativity. For a true phase transition of first order n goes to infinity. For the calculated curves in the figure the calorimetric transition enthalpy ΔH_{cal} was assumed to be 100 kcal mol^{-1}, which is characteristic for the denaturation transition of lysozyme (Figure 4.5). Then n was varied between 1, 2.5, 5, and 10. The transition peak narrows with increasing n, while the area under the curve, representing ΔH_{cal}, stays constant.

Figure 4.6 shows the curve from Figure 4.4 in the normalized form. In addition, the calculated curve assuming a two-state transition is shown. Evidently, this simple model already describes this transition quite well. The values for T_m and ΔH_{cal} were 74 °C and 100 kcal mol^{-1}, respectively. A better fit can be obtained using a nonlinear fitting procedure with n, ΔH_{cal}, and T_m being adjustable parameters. The red curve was obtained from this fitting and yielded $\Delta H_{cal} = 91.5$ kcal mol^{-1} and $n = 1.3$.

Not all proteins show an unfolding transition that can be described by a two-state process. In many cases specific domains of the protein unfold separately at different temperatures. This leads to DSC curves with several overlapping peaks. As an example Figure 4.7 shows the unfolding transition of an immunoglobulin IgG.

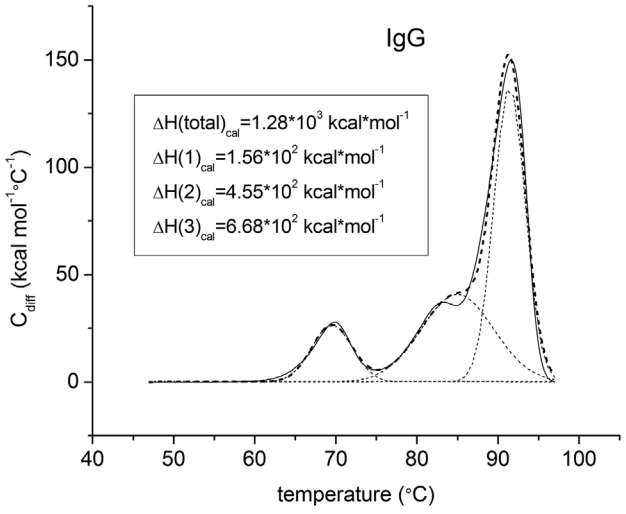


Figure 4.7 Normalized DSC curve (black) for the unfolding of an immunoglobulin showing several steps in the unfolding process. The simulated curve (bold dashed) is not based on a specific unfolding model but was calculated using three different Gauss functions to separate the transitions and to enable the determination of the three different transition enthalpies.

Integration of the separate DSC peaks is performed by using three different Gauss functions to separate the different contributions. Using this procedure the percentage of the three transitions in relation to the total transition enthalpy can be evaluated relatively simply. This procedure is implemented in many applications delivered with the DSC instruments by the manufacturer. The simplification using Gauss functions is performed because it is problematic to use the assumption of three different two-state transitions being present for the unfolding process, as there is a coupling of these three transitions.

The commercial DSC instruments available today are usually also capable of recording cooling curves, though with less sensitivity due to the inherent setup of a temperature gradient between the adiabatic shield and the cells. Using these instruments in the cooling mode makes it possible to test for the reversibility of the transitions and for hysteresis effects. For proteins, complete reversibility of the unfolding transition is rarely observed. This is evident from the different ΔH_{cal} values observed for unfolding and subsequent refolding and also from differences between the first and second heating curve. In many cases the second heating curve shows a shifted T_m and a lower transition enthalpy ΔH_{cal}. In many cases, heating a protein to temperatures above the unfolding transition can lead to irreversible aggregation processes. In the DSC curves this is discernible by sudden shifts in the baseline at temperatures above the transition to lower values, as these aggregation processes are in many cases connected with an exothermic heat. Phase transitions in lipid membranes, however, are in most cases completely reversible, showing only a small hysteresis upon cooling.

4.2 Isothermal Titration Calorimetry (ITC)

Like differential scanning calorimetry (DSC), isothermal titration calorimetry (ITC) is a widely used analytical method for the determination of binding constants and heats of binding of ligands bound to biopolymers and lipid membranes. Advances in technology in the last two decades have greatly improved the sensitivity of the method so that it is now widely established. The commercially available instruments have reached a high sensitivity so that heat effects down to 0.1 µcal (ca 0.4 µJ) can be measured. Binding constants ranging from 10^2 to 10^9 M^{-1} can be reliably determined using ITC. ITC has the advantage that it is a probe-free method and that it uses, like DSC, the differential principle, that is, changes in the binding of the ligand are detected by the heat being taken up or evolved during the titration. This differential method is generally more sensitive than methods in which the equilibrium concentrations of the bound and free ligand are determined.

The ITC instruments available today were derived from existing DSC instruments. The instruments work in the quasi-isothermal mode. The cells are heated with a constant but very low power leading to a slight increase in temperature over very long measuring times. This temperature increase is, however, negligible. The heating is necessary to enable a power compensation of exothermic as well as endothermic heat effects occurring during a titration run. The needle of a syringe equipped at the end with a small paddle is inserted into the sample cell. The needle or the complete syringe can be rotated by a motor for efficient stirring of the solution after injection of small amounts of solution (1–25 µl) into the sample cell. The reaction heat occurring after injection leads to a temperature difference between the cells which is detected. The electric signal is used to change the heating power of the sample cell. This differential power signal is recorded and displayed on the computer screen. Integration of the peaks yields the evolved or used heat during the reaction. The injections can be repeated until the total amount of solution in the syringe is used up or until no additional heat signals are observed. Figure 4.8 shows a scheme of an ITC instrument.

4.2.1 Ligand Binding to Proteins

Figure 4.9 shows the heat of reaction occurring due to the binding of a surfactant (Tween 80®) to bovine serum albumin (BSA). Usually, for the first injection a very small volume of only 1 µl is used to displace possible air bubbles captured in the injection needle. The subsequent

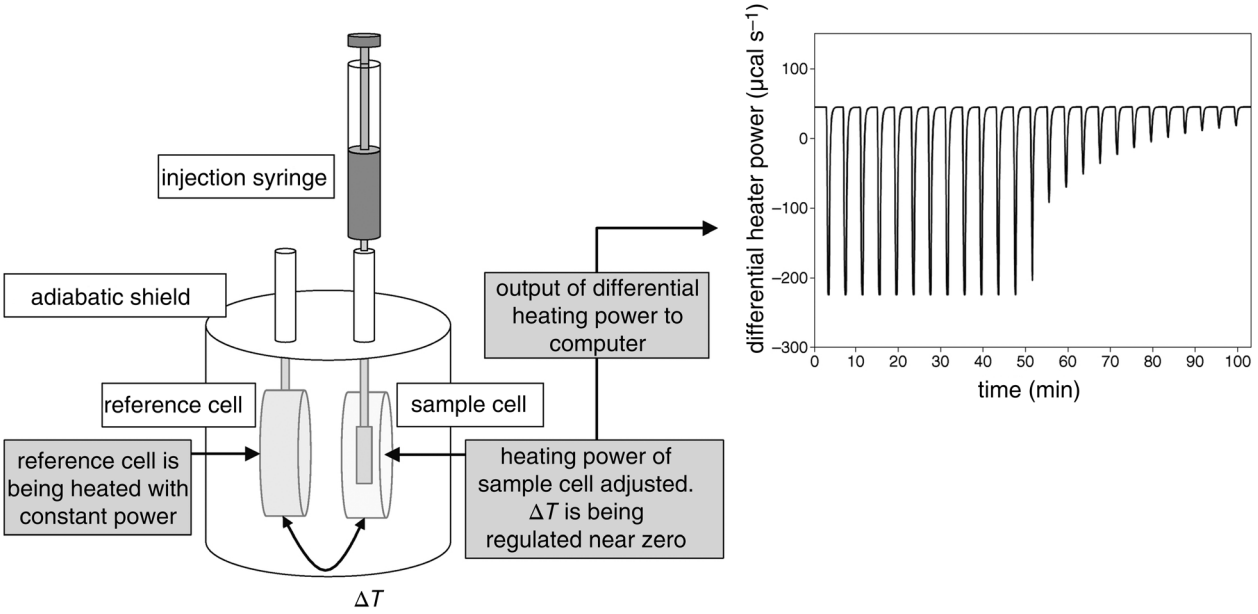


Figure 4.8 Scheme of an isothermal titration calorimeter (ITC). The principal construction is almost identical to a DSC instrument. An additional device is the motor driven injection syringe inserted into the sample cell. It is equipped with paddle at the end of the needle and is rotated by a motor to achieve stirring of the solution after injection of the ligand solution. The schematic registered curve shown in the graph was created by the injection of small amounts of solution from the syringe. Source: After Freyer and Lewis (2008) with permission. © 2008 Elsevier.

injections consist of 10 µl each. The heat of reaction decreases with increasing number of injections. The last observable heat effects are very small and are usually due to heats of dilution of the solution containing the ligand into the sample cell.

For analysis, the areas under the heat peaks are determined by integration and the values are normalized by considering the concentration and volume of injectant. The final output is the heat of reaction in energy per mole of injectant. The available commercial programs delivered by the manufacturer usually contain these evaluation procedures.

The experimentally determined values for the heat of reaction can also be simulated using different models for the binding of a ligand. For the simplest case of binding of only one ligand to a macromolecule, analysis of the titration curves is easy. The binding can be described by the law of mass action with L being the ligand, M the macromolecule, and ML the complex

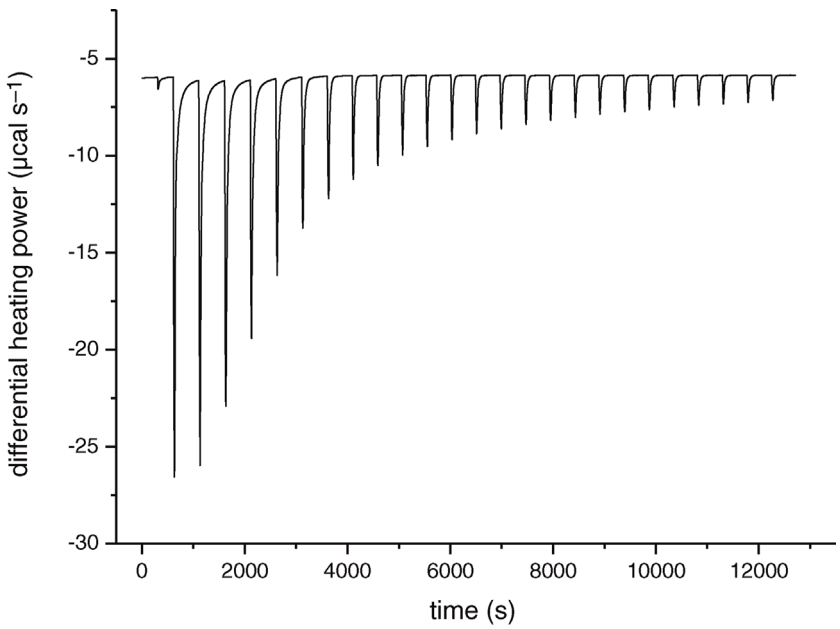


Figure 4.9 ITC curve for binding of Tween 80® to bovine serum albumin (BSA). A Tween 80® solution was titrated with 10 µl injections into the BSA solution in the sample cell. Source: Hoffmann, C., Blume, A., Miller, I., and Garidel, P. (2009) Insights into protein–polysorbate interactions analysed by means of isothermal titration and differential scanning calorimetry. *Eur. Biophys. J.*, **38,** 557–568.

between L and M (4.10). The binding constant K is then given by Equation 4.11, where K is expressed in units of mol l^{-1}:

$$L + M \rightleftharpoons ML \tag{4.10}$$

$$K = \frac{[ML]}{[L_f][M_f]} = \frac{\theta}{(1-\theta)[L_f]} \tag{4.11}$$

where [ML] is now the concentration of the complex, $[M_f]$ is the concentration of the free macromolecule, and $[L_f]$ is the concentration of the free ligand; θ is the degree of binding. It represent the ratio $[ML]/[M_t]$, with $[M_t]$ being the total concentration of the macromolecule.

The concentration of the free ligand and the free macromolecule are related to the total concentrations of the macromolecule or the ligand by:

$$[L_f] = [L_t] - [ML] \tag{4.12a}$$

$$[M_f] = [M_t] - [ML] \tag{4.12b}$$

Insertion of Equation 4.11 leads to:

$$K = \frac{[ML]}{([L_t] - [ML])([M_t] - [ML])} \tag{4.13}$$

Solving for the concentration [ML] leads to a quadratic equation:

$$[ML] = \tfrac{1}{2}\left([M_t] + [L_t] + \frac{1}{K}\right) - \sqrt{\frac{1}{4}\left([M_t] + [L_t] + \frac{1}{K}\right)^2 - [M_t][L_t]} \tag{4.14}$$

This equation describes the concentration of the complex ML as a function of the total concentration of L or M. In the experiments the ligand L contained in the syringe is titrated into the solution of the macromolecule M in the sample cell. Due to the injection of L the total concentration of L in the sample cell is changed by $\Delta[L_t]$, leading to a change in the concentration of the complex of Δ[ML]. This means that Equation 4.14 has to be differentiated with respect to $[L_t]$. This leads to:

$$\frac{\Delta[ML]}{\Delta[L_t]} = \frac{1}{2} - \frac{2\left([M_t] + [L_t] + \frac{1}{K}\right) - [M_t]}{2\sqrt{\frac{1}{4}\left([M_t] + [L_t] + \frac{1}{K}\right)^2 - [M_t][L_t]}} \tag{4.15}$$

The observed heat of reaction is then Equation 4.15 multiplied by the molar heat of binding for the ligand, ΔH_{bind}:

$$Q = \frac{\Delta[ML]}{\Delta[L_t]}\Delta H_{bind} + \Delta H_{dil} \tag{4.16}$$

The term ΔH_{dil} is added to account for heats of dilution of the ligand solution injected into the sample cell. These heats of dilution can be separately determined by injection of the ligand solution into a sample cell filled with buffer only or by evaluation of the last injection peaks when the reaction is over as shown in Figure 4.9.

It is also possible to perform the experiment by reversing the solution in the syringe and in the cell, that is, by titrating the solution of the macromolecule into the solution of the ligand. Equation 4.14 has then to be differentiated with respect to the total concentration of the macromolecule, giving the following equation:

$$\frac{\Delta[ML]}{\Delta[M_t]} = \frac{1}{2} - \frac{2\left([M_t] + [L_t] + \frac{1}{K}\right) - [L_t]}{2\sqrt{\frac{1}{4}\left([M_t] + [L_t] + \frac{1}{K}\right)^2 - [M_t][L_t]}} \tag{4.17}$$

Completely equivalent to Equation 4.16, the heat of reaction is obtained by the following equation, where ΔH_{dil} is now the heat of dilution of the macromolecule when injected into the ligand solution:

$$Q = \frac{\Delta[ML]}{\Delta[M_t]}\Delta H_{bind} + \Delta H_{dil} \tag{4.18}$$

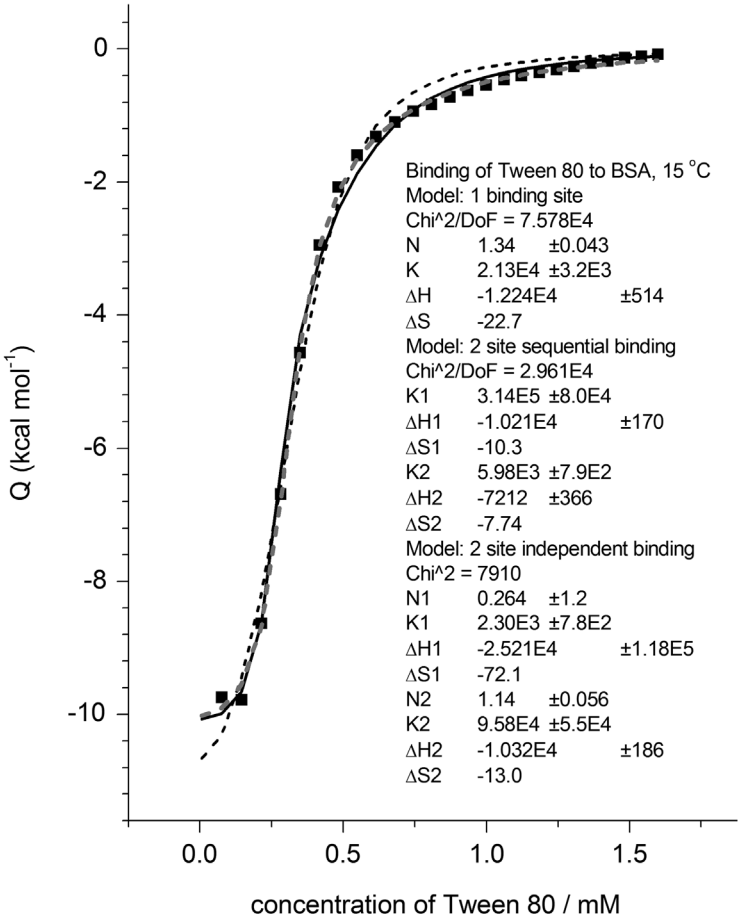


Figure 4.10 Normalized titration curve for the binding of Tween 80® to BSA (black squares). The dashed black curve is a simulation using a simple binding model with one binding site. The full black curve was calculated using a model with sequential binding sites and different binding constants. An even better agreement with the experimental curve can be achieved using a binding model with two independent binding sites (dashed grey curve). However, the obtained binding parameters are physically not meaningful (see text for explanation).

Figure 4.10 shows a normalized binding curve and simulated curves based on different binding models.

The dashed red curve in the figure was obtained by simulation with a one-site binding model. There are significant deviations between the simulated and experimental curves. In particular, the stoichiometry factor $N = 1.34$ shows that possibly several binding sites are present. Two other simulations were then performed using different binding models. The second model uses a sequential binding mechanism with two steps and two different binding constants:

$$\mathrm{L} + \mathrm{M} \rightleftharpoons \mathrm{ML} \tag{4.19a}$$

$$\mathrm{ML} + \mathrm{L} \rightleftharpoons \mathrm{ML_2} \tag{4.19b}$$

The resulting binding constants K_1 and K_2 are:

$$K_1 = \frac{[\mathrm{ML}]}{[\mathrm{L_f}][\mathrm{M_f}]} \tag{4.20a}$$

$$K_2 = \frac{[\mathrm{ML_2}]}{[\mathrm{L_f}][\mathrm{ML}]} \tag{4.20b}$$

The relations between the different concentrations are:

$$[\mathrm{L_t}] = [\mathrm{L}] + [\mathrm{ML}] + 2[\mathrm{ML_2}] \tag{4.21a}$$

$$[\mathrm{M_t}] = [\mathrm{M}] + [\mathrm{ML}] + [\mathrm{ML_2}] \tag{4.21b}$$

In an analogous way as described above for the one-site binding model, the equations have to be transformed to obtain expressions for the derivatives $\Delta[\mathrm{ML}]/\Delta[\mathrm{L_t}]$ and $\Delta[\mathrm{ML_2}]/\Delta[L_t]$. Using

equations similar to Equation 4.16 containing the binding enthalpies $\Delta H_{1\text{bind}}$ and $\Delta H_{2\text{bind}}$ they are connected with the experimentally observed heat of reaction Q. The full red curve shown in Figure 4.10 was obtained by fitting the experimental curve with this model.

This latter curve shows a significantly better fit, which indicates that a second binding site exists with lower affinity.

The experimental curve can also be simulated with a model using two independent binding sites. The corresponding equations for the binding equilibria are:

$$L + M \rightleftharpoons ML_a \tag{4.22a}$$

$$L + M \rightleftharpoons ML_b \tag{4.22b}$$

The two binding constants K_a and K_b are defined as:

$$K_a = \frac{[ML_a]}{[L_f][M_f]}; \quad K_b = \frac{[ML_b]}{[L_f][M_f]} \tag{4.23}$$

Figure 4.10 also shows a black dashed curve calculated with this model. It gives an even better fit, as shown by the lower χ^2 values (sum of squared errors), but leads to parameters that are physically not meaningful, namely, a very low stoichiometry factor of 0.264 with an extremely large error as well as a value of $\Delta H_{1\text{bind}}$ for which the error is also larger than the absolute value of the heat of binding. This shows that simulations with the lowest value of the sum of squared errors χ^2 are not necessarily the best as they can lead to parameters that physically not meaningful.

4.2.2 Binding of Molecules to Membranes: Insertion and Peripheral Binding

The previous section described the binding of ligands to a macromolecule in solution. ITC is also widely used for binding studies of ligands to model membrane systems, such as lipid vesicles. Hydrophobic as well as amphiphilic molecules can partition into the hydrophobic region of a lipid bilayer. The difference between this partitioning and the binding of a ligand to a macromolecule is that no specific binding site exists in the bilayer membrane. However, the experimental approach to studying the binding or partitioning is similar to the experiments described above. The hydrophobic or amphiphilic molecule is either titrated to the suspension of lipid vesicles in the sample cell or vice versa, that is, lipid vesicles are titrated to the solution of the molecules in the sample cell. As mentioned above, no specific binding sites exist in the lipid membrane, but the molecules partition into the hydrophobic interior of the membrane or into the interface region between the headgroups and the hydrocarbon chains. This partitioning occurs under expansion of the membrane. An increasing number of incorporated molecules can lead to a change in the aggregation state, for instance, to the formation of micellar aggregates. As this change in aggregation state is in many cases unwanted, the concentration ratio of lipid to hydrophobic molecule has to be chosen such that an excess of membrane lipids is always present. This means that the effective molar ratio of lipid: ligand in the membrane must be kept larger than 10: 1. To calculate this effective ratio the binding constant to the membrane or partitioning constant of the molecule between water and membrane has to be known. This constant can in turn be determined from the ITC experiment. The equations describing the binding to a membrane are similar to the equation used for the simple binding described in Equation 4.10. A partitioning of the molecule between water and the membrane with a coefficient P is assumed. The term P is best expressed as the ratio of the concentration expressed in mole fraction units:

$$P = \frac{x_e}{x_W} \tag{4.24}$$

where x_e is the mole fraction of the molecule in the membrane and x_W is the mole fraction of the molecule dissolved in water:

$$x_e = \frac{D_e}{D_e + L} \tag{4.25}$$

$$x_W = \frac{D_W}{D_W + W} \tag{4.26}$$

D_e is then the molar concentration of the hydrophobic molecule in the bilayer and D_W in water. Both concentrations are normalized to the total volume of the sample. Here L is the lipid concentration and W the concentration of water (55.55 mol l^{-1}). The partition coefficient can thus be expressed as:

$$P = \frac{D_e(D_W + W)}{(D_e + L)D_W} \tag{4.27}$$

where $D_W = D_t - D_e$ (D_t is the total concentration) and $D_W + W \approx W$, as W is usually much higher than D_W. This leads to:

$$D_e = \frac{1}{2P}\left[P(D_t - L) - W + \sqrt{P^2(D_t + L)^2 - 2PW(D_t - L) + W^2}\right] \tag{4.28}$$

In an experiment where lipid vesicles are titrated into a solution of the hydrophobic molecules the change in D_e after addition of ΔL moles of lipid has to be calculated to simulate the titration curve:

$$\frac{\Delta D_e}{\Delta L} = -\frac{1}{2} + \frac{P(D_t + L) + W}{2\sqrt{P^2(D_t + L)^2 + 2PW(L - D_t) + W^2}} \tag{4.29}$$

The observed heat of reaction is then:

$$Q = \frac{\Delta D_e}{\Delta L}\Delta H_T + \Delta H_{dil} \tag{4.30}$$

with ΔH_T being the molar enthalpy of partitioning between water and the membrane and ΔH_{dil} the heat of dilution of the lipid vesicles into buffer. ΔH_T is now determined using a nonlinear fitting algorithm with ΔH_T and P as fitting parameters. ΔH_{dil} can be determined from a separate dilution experiment or from the last peaks with constant area in the ITC titration curve. Figure 4.11 shows an experiment where an amphiphilic molecule, the detergent octylglucoside (OG), partitions into lipid vesicles of the phospholipid dimyristoyl-phosphatidylcholine (DMPC).

In this experiment lipid vesicles (DMPC, $c = 50$ mM) were titrated into a solution of 3 mM octylglucoside, that is, a concentration below the critical micellization concentration of the detergent. Incorporation of the detergent in the membrane leads to exothermic heat effects that decrease with the number of injections. The normalized reaction heat as a function of lipid concentration can now be analyzed and the data can be fitted using the partitioning model described above. The resulting parameters for the partition coefficient and the transfer enthalpy of the detergent from water to the bilayer are $P = 3479$ and $\Delta H_T = -2.16$ kcal mol^{-1}.

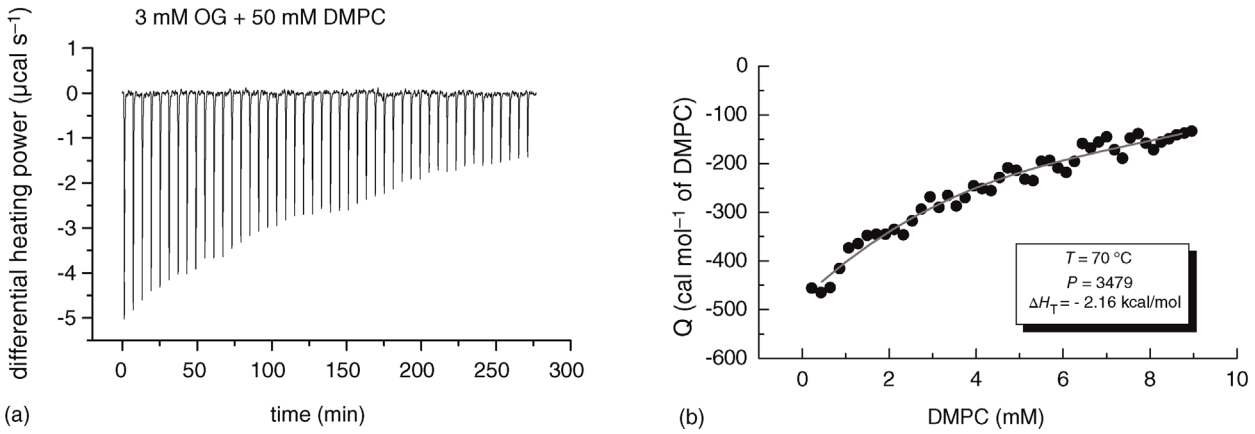


Figure 4.11 (a) Experimental ITC curve for the titration of a suspension of DMPC vesicles into an octylglucoside (OG) detergent solution in the cell. The heat effects are due to the partitioning of the detergent into the lipid bilayer. (b) Normalized ITC curve for the incorporation of OG into DMPC vesicles. The gray full line is a fit based on the partitioning model of OG between water and bilayer (see text). The values of the partition coefficient P and the transfer enthalpy ΔH_T are shown in the graph.

For a titration of the hydrophobic molecule to lipid vesicles the following equation can be derived:

$$\frac{\Delta D_e}{\Delta D_t} = +\frac{1}{2} + \frac{P(D_t + L) - W}{2\sqrt{P^2(D_t + L)^2 + 2PW(L - D_t) + W^2}} \tag{4.31}$$

The analysis presented above is based on the assumption that the partitioning is ideal, that is, that P does not depend on the effective concentration x_e of the molecule in the bilayer. This, however, is rarely the case, so that one has to deal with a x_e-dependent partition coefficient P. The simplest approach to deal with this problem is to introduce a non-ideality parameter ρ for the partitioning, leading to a concentration dependent P. This leads to the following equation:

$$P = P(x_e = 1)\exp\left[-\frac{\rho(1 - x_e)^2}{RT}\right] \tag{4.32}$$

where $P(x_e = 1)$ is the partition coefficient for ideal mixing in the bilayer. The analysis now becomes more complex and a simple fitting of the curves by nonlinear fitting algorithms is no longer possible. For a detailed description of the analysis procedure the reader is referred to the original articles (e.g., Keller, M., Kerth, A., Blume, A. (1997) Thermodynamics of interaction of octyl glucoside with phosphatidylcholine vesicles: partitioning and solubilization as studied by high sensitivity titration calorimetry. *Biochim. Biophys. Acta,* **1326**, 178–192) or to reviews (see Further Reading).

In the case of binding of molecules to lipid vesicles the binding can also occur to the surface of the vesicles, for instance, when the binding mode is by electrostatic attraction. In this case the analysis is much simpler, particularly if corrections due to build-up of an electrical double layer can be neglected. This is the case when the ionic strength of the solution is approximately equivalent to a physiological salt solution. As an example the binding of the cationic oligopeptide pentalysine (lys_5) to a negatively charged lipid vesicle composed of dimyristoyl-phosphatidyl-glycerol (DMPG) is shown. Pentalysine has five positively charged side groups and thus binds by attractive electrostatic interactions to the negatively charged vesicle surface. With ITC, heat effects due to the binding reaction can be observed. These heat effects are mainly due to changes in hydration at the vesicle surface and the liberation of sodium counterions into the bulk solution. Figure 4.12 shows a titration curve for pentalysine binding to fluid DMPG vesicles.

The titration curve has a sigmoidal shape, indicating saturation above a stoichiometric ratio of 5, as expected for stoichiometric binding of one pentalysine to five DMPG molecules which are singly charged. Obviously, all lipid headgroups are accessible for binding. This means that pentalysine can cross the vesicle membrane and can reach the interior surface of the vesicle. An

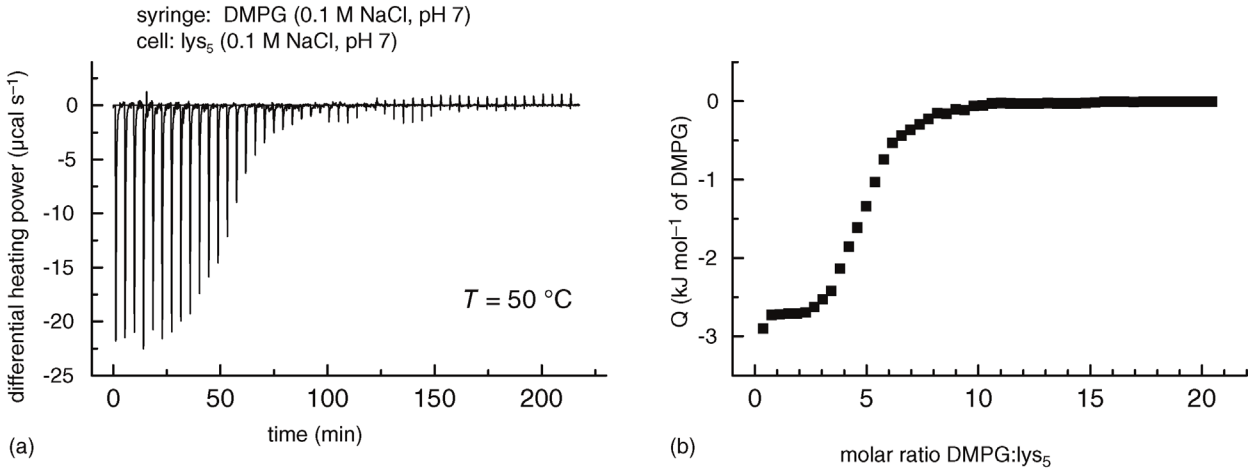


Figure 4.12 (a) Experimental ITC curve for the titration of lipid vesicles of DMPG in the syringe to a solution of pentalysine (lys_5) in the cell. The heat effects are caused by electrostatic binding of the cationic lys_5 to the negatively charged DMPG vesicle surface. (b) Normalized ITC titration curve for the binding of lys_5 to negatively charged DMPG membranes.

intermediate pore formation, the formation of a defect, or the complete destruction of the vesicle has to be assumed so that the headgroups in the interior become accessible. Temperature dependent ITC experiments have shown that at lower temperature outside binding only of pentalysine can occur so that the assumption of transient defects in fluid vesicles is the most likely explanation. For analysis of the binding process a model analogous to Equation 4.15 can be used with a modified expression for the binding constant:

$$lys_5 + PG_5 \rightleftharpoons (lys_5 \cdot PG_5) \tag{4.33}$$

$$K = \frac{[(lys_5 \cdot PG_5)]}{[lys_5][PG_5]} \tag{4.34}$$

Fitting of the titration curve shown in Figure 4.12 leads to a binding constant K of ca. $2{\times}10^5\ M^{-1}$ and a binding enthalpy ΔH of -28 kJ mol^{-1} based on lys_5. The great advantage of ITC over other methods for the determination of binding constants is the fact that two thermodynamic quantities are determined from one and the same experiment, namely, the binding constant K and the binding enthalpy ΔH^0_R. When spectroscopic methods are used, the binding enthalpy can only be determined by temperature dependent measurements of the binding constants using the van't Hoff equation for the determination of the binding enthalpy. The binding constant K is related to the standard free energy of binding ΔG^0_R by:

$$\Delta G^\circ_R = -RT \ln K \tag{4.35}$$

and the relation between ΔG°_R and ΔH°_R is given by the Gibbs–Helmholtz equation:

$$\Delta G^\circ_R = \Delta H^\circ_R - T\Delta S^\circ_R \tag{4.36}$$

In this way also the standard entropy of binding ΔS°_R can be calculated. As mentioned above, when using other methods these thermodynamic quantities can only be determined by measurements of the temperature dependence of the binding constant K, which have to be very precise and are thus time-consuming. Interpretation of the thermodynamic parameters using molecular models is not a simple task and compensating effects occurring in the heat of binding can lead to different possible explanations in terms of molecular models. In the case of a strong temperature dependence of the heat of binding ΔH°_R the change in heat capacity ΔC°_{pR} between products and educts is responsible and can be determined from the slope of a ΔH°_R versus temperature plot. For experiments with macromolecules or lipid vesicles in aqueous solution changes in hydration of the products compared to the educts play a major role in determining ΔC°_{pR}. In particular, changes in hydrophobic hydration, that is, the exposure of hydrophobic groups, leads to a positive contribution to the ΔC°_{pR} values, so that estimates about the exposure of hydrophobic surfaces is possible.

4.3 Pressure Perturbation Calorimetry (PPC)

Pressure perturbation calorimetry (PPC) is a new calorimetric method introduced in recent years, which has attracted attention. For these experiments a standard high sensitivity DSC instrument with fixed cells can be used. The cells are heated very slowly with low heating power and a pressure jump of 4–5 bar is applied to the solutions in both cells. This leads to a heat signal due to the compression or expansion of the solutions. The exchanged heat due to the pressure jump is related to the coefficient of thermal expansion of the solution. Comparing the coefficients of thermal expansion of a buffer solution to those of a protein solution in buffer yields the coefficient of thermal expansion of the dissolved protein at defined temperatures. Integration of the coefficient of thermal expansion over the temperature range of a thermal denaturation of a protein in solution leads to the apparent volume change of the protein occurring upon unfolding. PPC is not a routine method, as the experiments are very time consuming, because a wide temperature range for the data is needed. In addition, interpretation of the data is complicated and in many cases ambiguous. The reason behind these problems lies in the fact that the coefficients of thermal expansion and the volume change during unfolding are not only due to the protein itself but also to the hydration of the protein and its changes during unfolding. Review articles cited in the Further Reading section provide further information.

Further Reading

Bastos, M. (ed.) (2016) *Biocalorimetry: Foundations and Contemporary Approaches*, CRC Press: Boca Raton, FL., USA.

Blume, A. (2016) Temperature induced and isothermal phase transitions of pure and mixed lipid bilayer membranes studied by DSC and ITC, in *Biocalorimetry: Foundations and Contemporary Approaches* (ed. M. Bastos), CRC Press, Boca Raton, FL, pp. 103–136.

Blume, A. (1988) Applications of calorimetry to lipid model membranes, in *Physical Properties of Biological Membranes and their Functional Implications* (ed C. H. Hidalgo), Plenum Press, New York, pp. 71–121.

Blume, A. and Garidel, P. (1999) Lipid model membranes and biomembranes, in *Handbook of Thermal Analysis and Calorimetry, Volume 4: From Macromolecules to Man* (ed. R. B. Kemp), Elsevier Press, Amsterdam, pp. 109–173.

Danforth, R., Krakauer, H., and Sturtevant, J.M. (1967) Differential calorimetry of thermally induced processes in solution. *Rev. Sci. Instrum.*, **38**, 484–487.

Freiburger, L.A., Auclair, K., and Mittermaier, A.K. (2009) Elucidating protein binding mechanisms by variable-c ITC. *ChemBioChem*, **10**, 2871–2873.

Freyer, M.W. and Lewis, E.A. (2008) Isothermal titration calorimetry: experimental design, data analysis, and probing macromolecule/ligand binding and kinetic interactions. *Methods Cell. Biol.*, **84**, 79–113.

Hemminger, W. and Höhne, G. (1984) *Calorimetry: Fundamentals and Practice*, Verlag Chemie, Weinheim.

Heerklotz, H. and Blume, A. (2012) Detergent interactions with lipid bilayers, in *Comprehensive Biophysics* (ed E. G. Egelman), Academic Press, Oxford, vol. 5, pp. 63–91.

Murphy, J., Knutson, K., and Hinderliter, A. (2009) Protein–lipid interactions. *Methods Enzymol*. **466**, 431–453.

Plotnikov, V.V., Brandts, J.M., Lin, L.-N., and Brandts, J.F. (1997) A new ultrasensitive scanning calorimeter. *Anal. Biochem.*, **250**, 237–244.

Privalov, P.L., Ptitsyn, O.B., and Birshtei, T.M. (1969) Determination of stability of the DNA double helix in an aqueous medium. *Biopolymers*, **8**, 559–571.

Schweiker, K.L. and Makhatadze, G.I. (2009) Use of pressure perturbation calorimetry to characterize the volumetric properties of proteins. *Methods Enzymol.*, **466**, 527–547.

Spink, C.H. (2008) Differential scanning calorimetry. *Methods Cell. Biol.*, **84**, 115–141.

Wiseman, T., Williston, S., Brandts, J.F., and Lin, L.N. (1989) Rapid measurement of binding constants and heats of binding using a new titration calorimeter. *Anal. Biochem.*, **179**, 131–137.

Immunological Techniques

5

Reinhold P. Linke

Reference Center of Amyloid Diseases anYmed, Innovation Center of Biotechnology (IZB), Vinzenz-Schüpfer-Straße 20a, 81475 Munich, Germany

Thanks to their high specificity in the recognition of antigens, especially those that differ only slightly, antibodies have won increasing importance in bioanalytics as well as in therapeutics.

As the induction and synthesis of antibodies *in vivo* is very complex, some terminology necessary for an understanding of the synthesis, structure, and function of antibodies will be explained. This chapter considers mainly the application of antibodies as analytical reagents in a wide range of immunoassays for innumerable applications, extending to modified and completely synthetic immunotherapeutics.

5.1 Antibodies

5.1.1 Antibodies and Immune Defense

The concept of antibody refers to humoral and cellular multifunctional glycoproteins from vertebrates that essentially serve as defense against microorganisms and viruses. The antibodies bind microorganisms by means of the antigen–antibody reaction, which functions according to the "lock and key" principle. The chemical complexity of antibodies has increased in the course of evolution, especially since the land colonization by vertebrates in the Devonian period and the appearance of amphibians. In mammals, both the antibody system, which serves for humoral defense, and the chemically homologous major histocompatibility complex (MHC) that provides cellular immunity belong to the protein families with the greatest chemical diversity attained during evolution. Antibodies are found only in chordates, beginning with the lampreys (Petromyzontiformes).

Antibodies, like all members of the antibody super family, essentially consist of receptors. Antibodies and the MHC serve primarily for the distinction between individual body substances ("self") and "non-self" (foreign) body substances for the purpose of eliminating foreign material to maintain the integrity of the organism without bringing it into danger itself. This requires a high degree of distinguishing power between self and non-self and a precise elimination mechanism.

This extremely precise distinguishing ability by the innate and in particular the adaptive immune system has enabled the evolution of vertebrates.

The huge number of various microorganism species, which could take over all imaginable environments, with their very short generation times and great ability to change through genetic mutations, were countered by the similarly changeable adaptive immune system of the long-living vertebrates. This immune system is in regard to the possible specificities so widely dimensioned that it eliminated a great number of newly appearing microorganisms as well as practically all synthetically produced laboratory substances that were never anticipated during

Immunity, humoral-cellular The protection of vertebrates against microbial and viral invasion is based on two pillars: (1) on the soluble antibodies in the extracellular fluids (blood plasma, secretions, and lymph) by means of which protection is transferred to another individual (passive immunity, introduced by Emil von Behring, awarded the Nobel Prize in Physiology or Medicine 1901); (2) on various immune cells of the hematologic system, which carry on their surface cell-associated antibodies and chemically related recognition structures on invading foreign organisms. This protection is also transferable, but only via immune cells.

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

evolution. Thus, the immune system is so constructed that immunotherapeutics can be produced against practically any specificity.

5.1.2 Antibodies as Reagents

An antigen–antibody reaction, by means of which foreign organisms can be recognized and eliminated *in vivo,* has been reproduced *in vitro* in simple experiments since the end of the nineteenth century. Since then, knowledge concerning humoral and cellular immune defense and their biochemical mechanisms, especially the discriminating properties of the antibody, has increased tremendously. Karl Landsteiner (Nobel Prize in Physiology or Medicine 1930) demonstrated the exorbitantly huge number of various specificities and showed that antibodies can distinguish various conformations of antigens as wells as positional isomers such as *ortho*- and *meta*-nitrophenyls.

A vast number of experiments have led to laboratory tests that, by means of the antigen–antibody reaction, enable identification, quantification, and isolation of all imaginable antigens or antibodies.

By means of antibodies it is, for example, possible to exactly identify a great number of proteins and thus to smooth the way to an explanation of their functionality.

Antibodies have penetrated into all branches of biosciences as ideal reagents since the identification of a great number of proteins and other substances require only a single identification principle, namely, the antigen–antibody reaction. Since immunoassays are relatively easily handled and dependable they have in general proved suitable for automation.

This chapter illustrates some important principles of the antigen–antibody reaction that nature has developed for immune defense and which scientists and practitioners in medicine and industry in the form of manifold test methods now use for analytic, diagnostic, therapeutic, and preparative purposes. After a short presentation of the structure and characteristics of antibodies and antigens as well as the antigen–antibody reaction, typical immunological techniques will be treated as well as the production of antibodies, especially concerning the modern *engineering* of antibodies leading to the production of completely synthetic antibodies.

5.1.3 Properties of Antibodies

Isotype Homologs of closely related proteins that are coded in duplicate structure genes (at different gene loci!) and are all co-expressed in an individual. Example: the immunoglobulins A, D, E, G and M, the IgG-subclasses 1–4 or the two immunoglobulin light chains and with their sub-groups (see Allotype).

In mammalian organisms five classes of antibodies are distinguished: immunoglobulins A, D, E, G, and M, abbreviated IgA, IgD, and so on, which all exist simultaneously in the plasma of an individual organism and are therefore named isotypes. It must nevertheless be stated that the composition of immunoglobulins (Ig) in plasma, exocrine secretions, and in tissues differ and that various immunoglobulin classes are induced according to their immunization type and production location. The functions of antibodies after antigen interaction are quite manifold. A number of these functions are applicable in the form of standard tests, some of which are described here in detail. Antibodies are divalent (e.g., IgG), tetravalent (IgA), up to decavalent (IgM) and crosslinked natural multivalent antigens (Figure 5.6 below). The growing immune complexes lose their solubility when forming aggregates and fall out of solution as white flocks in the process called precipitation (Section 5.3.2).

Clone, **clonal**, **monoclonal** If a cell multiplies without further differentiation by continuous cell division one speaks of *clonal growth*. The result is the making of a *cell clone*. When every cell of an antibody-making clone synthesizes the same antibody, it is called *monoclonal.*

IgM and IgG bind and also activate complement after antigen-interaction and finally lead through this to vicinal cell membrane lysis (including bacteria, viruses, fungi, and parasites) and to activation of other functions (Section 5.4). The cell adhesion by an antibody leads to secondary reactions like, for example, phagocytosis and cell-mediated cytotoxicity (Section 5.5).

Although vertebrate IgA is the most abundant immunoglobulin, its employment as analytical reagent is restricted to only special cases. It is found mostly on mucus membrane surfaces of the digestive tract, the lungs, in exocrine gland secretions, as well as in blood plasma, and is in general less easily accessible. IgA in contrast to IgG (see below) constitutes only 15% of the

plasma immunoglobulins and is heterogeneous in size (monomer 160 kDa, dimer 390 kDa, secretory dimer 385 kDa, higher polymers).

IgM is a decavalent pentameric immunoglobulin (970 kDa) that constitutes about 7% of the plasma immunoglobulins. Because of its size it is less soluble and precipitates reversibly from normal serum when diluted tenfold with double-distilled water. Its adsorption to all possible surfaces can trigger manifold unspecific interactions. Therefore, the application of IgM, like IgA, is restricted to certain questions.

IgD (180 kDa) and IgE (190 kDa) are antibodies that are found mainly on lymphocytes and mast-cell surface membranes. They appear only as traces (0.5% and 0.002%) in the plasma immunoglobulins. IgD has no significance as an analytical antibody. IgE, in contrast, mediates allergic reactions and therefore plays a central role in allergies.

Fc-Receptor Binding structures on the cell surfaces of, for example, monocytes and lymphocytes that bind to the Fc-part of an antibody (Figure 5.1). The Fc-receptor binds (via antigen or aggregation) crosslinked antibodies. As a result, this bonding has a series of important defense functions, for example, phagocytosis and antibody-mediated cytotoxicity.

IgG dominates in terms of amount in plasma and extracellular space (but not on the mucus membrane surfaces) all other immunoglobulins (about 75% of all immunoglobulins) and represents *the* analytical antibody of immunochemistry. IgG antibodies are inducible in animals by immunization and are easy to obtain by means of blood vessel puncture (Section 5.7). In addition, the aiding of adjuvant-induced antibodies of experimental animals and also the corresponding monoclonal antibodies (Section 5.7) are mostly of IgG-type.

As a monomer of size 150 kDa, IgG is smaller than the representatives of the other immunoglobulin classes. There are four isotypes of IgG ($IgG_{1,2,3}$ and IgG_4) that are differently bound to bacterial Fc-receptors, for example, protein A from *Staphylococcus aureus*. Protein A is used in various immunoassays (Section 5.3). The four IgG isotopes carry a train of allotypic (inheritable according to the laws of Gregor Mendel) mutations that are called Gm-factors and are referred to in genetic studies.

Allotype One or several amino acid replacements, coded in one of the allelic genes (at one gene position!). Allotypes appear only in some individuals of a population; they differ from sporadic amino-acid replacements by the fact that they have increased in the total population to over 1%. Example: blood group substances or the immunoglobulin-Gm-factors (see Isotype).

IgG is easily soluble and remains active even in reduced salt concentration. IgG-antibodies react in a pH range of about 4–9.5. IgG can be stored in sterile serum at 4 °C from months up to several years. The functional half-life of IgG in serum after shock-freezing (in CO_2 dry ice/alcohol or even better in liquid nitrogen) can extend to decades when stored at −20 °C or better at −80 °C. Antibodies can be reversibly denatured, so that an antigen–antibody bonding can be released again without the antibody suffering too much damage (Box 5.1).

This property allows isolation of antigens or antibodies from complex protein mixtures by means of affinity chromatography when one of the bonding partners is immobilized (Section 5.3.3).

Affinity Chromatography, Section 10.4.8

In addition, for IgG, its antibody activity can be coupled covalently to a train of amplifying substances. These amplifying substances, like fluorochromes, enzymes, and radionuclides, serve to increase the sensitivity for identification and quantitation of bound antibody in interaction tests (Section 5.3.3).

Box 5.1 Reversible separation of antigen–antibody-complexes *in vitro*.

Acidic conditions	The optimal pH is 2.6; in the case of high affinity antibodies more extreme conditions, for example, a short period at 4 °C at pH 1.8 – with, however, more severe antibody damage – may be necessary for the recovery of a bound antigen
Alkaline conditions	The optimum pH is 11.2; with the application of harsher conditions the same is valid as with acid conditions. Nevertheless, the antibodies are more damaged by alkaline than acid conditions
Chaotropic ions	Cl^-, I^-, Br^-, SCN^-; typical eluents are 3M $MgCl_2$, 1–3M NaSCN
Epitopes	Higher concentrations of competing free antigens, of synthetic peptides, isolated epitopes, or immunodominant epitope segments, for example, from free haptens (Figure 5.2b, c)
Higher temperature	Nowadays considered unusable

5.1.4 Functional Structure of IgG

Firstly it must be noted that all that is said in the following relates essentially to the IgG_1-antibody. In 1967 Gerald Edelman (Nobel Prize in Physiology or Medicine 1972) described the first amino acid sequence and the complete covalent structure of an IgG_1-antibody. In Figure 5.1, an IgG_1-antibody is illustrated schematically.

The immunoglobulin polypeptide chains each consist of a series of homologous repeats of a single ancient primary immunoglobulin unit of 12 kDa, which carries a disulfide bridge that forms a loop of 60 amino acids.

The heavy chain has four homologous regions and the light chain has two, from which each corresponding to the primary Ig-chain is stabilized by its own disulfide bridge. They are connected by inter-chain disulfide bonds. The homologous regions of different chains are folded into domains, from which two relatively independent functional units (modules) lie together. For example, the N-terminal domains of the heavy and light chain come together to a variable module with the combining site formed by both chains (Section 5.1.5). Other modules carry out other functions (Figure 5.1).

The extremely compact modules can be enzymatically separated at the enzyme-sensitive bonds of polypeptide stretches between the modules. The fragments so obtained can be studied individually regarding their functions as well as by being employed in manifold tests. The gene-technical production of these segments in bacteria for special investigation or usage is also possible (Section 5.7.2). If, occasionally, a monovalency of an antibody is required, the half-molecules can be separated by mild reduction of the disulfide bridges between the H-chains, or the monovalent Fab-fragment can be obtained by enzymatic removal of the Fc-fragment and the *hinge*-region using papain. Pepsin cleavage produces the bivalent $F(ab')_2$-fragment with the intact *hinge*-region; consequently, the influence of a monovalent combining site and the effect of divalency without additional Fc-aggregation, or at the same time the function of the *hinge*, can be studied. In Figure 5.1 other typical fragments of IgG are shown and the functions of the IgG-domains are described. Section 5.7.2 gives a short description of how antibody *engineering* has led to greatly increased fields of application.

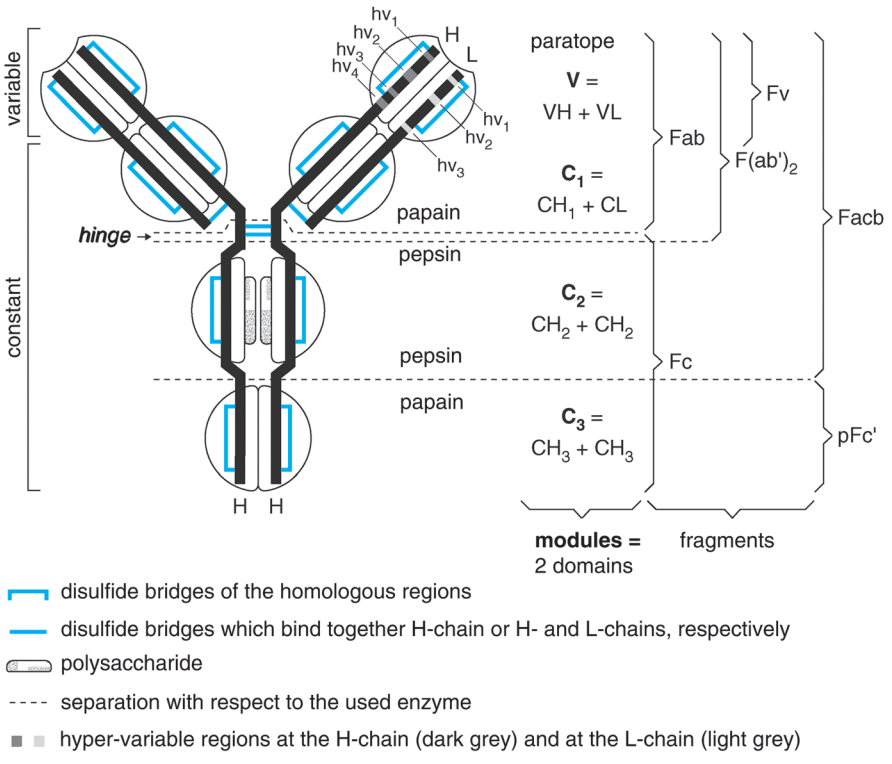


Figure 5.1 Schematic structure and function of an IgG molecule. An IgG_1 molecule consists of two identical heavy (H, 50 kDa) and two identical light (L, 24 kDa) polypeptides that are covalently bound together by means of disulfide bridges. Thus, two covalently bound, functionally identical units arise (2(H + L) each of which is 75 kDa), providing the most important bivalence necessary for functionality. These monovalent units are bound together at the hinge region via several disulfide bridges. This organization enables the angle between the Fab-fragment to vary greatly (for bridging epitopes and better epitope fitting). The modules are depicted as globular units, each consisting of two domains. The immunoglobulin molecule consists of one variable N-terminal module (V) and three following constant modules (C). The V (VH + VL) is the paratope carrying module with the seven hyper-variable regions at the H-chain and at the L-chain, of which six (see below) constitute the contact with the epitope of the antigen; C1 (CH_1 + CL_1) allows the interaction with complement component C_{4b}; C_2 (CH_2 + CH_2) allows complement activation and determines the metabolic degradation and protein-A interaction (together with C_3), C_2 contains a structural polysaccharide component. The C_3 module (CH_3 + CH_3) enables, together with C_2, attachment to the cell surface and induces cellular immune response via the Fc-receptor. Separation of the various modules after limited digestion with papain and pepsin is shown.

5.1.5 Antigen Interaction at the Combining Site

The combining site (paratope) is located on the variable module (V), which consists of the variable domains of the light and heavy immunoglobulin chains (VL and VH). The H-chain contains four hyper-variable regions (hv_1, hv_2, hv_3, and hv_4) and the L-chain has three (hv_1, hv_2, and hv_3). The close epitope contact is made by the H-chain's hv_1, hv_2, and hv_4 and all three hv regions of the L-chain.

These regions are therefore also known as complementarity determining regions (CDRs) and the stabilizing regions located in between are the framework determining regions (FDRs). The contribution of the H-chain to specificity is, at about 70%, greater than that of the L-chain, approx. 30%, so that the isolated H-chain retains antibody activity better than the corresponding isolated L-chain (Figure 5.2). The L-chain often loses antibody activity after separation from the H-chain.

A renewed combining of heavy and light chains can restore the antibody specificity and affinity (see below). Recombination with a different light chain leads, however, to a different antibody specificity. The size of the combining site can be measured with smaller antigens made of various numbers of amino acids. Hexamers to octamers show the maximal binding to the paratopes, comparable with that of intact immunogen. From these data a contact area of about 3 nm diameter can be derived for the combining site. This area, the paratope, can house a train of

Paratope By paratope (combining site) is meant the antibody-segment that interacts with the antigen. The antigen region bound to the paratope is called the *epitope*.

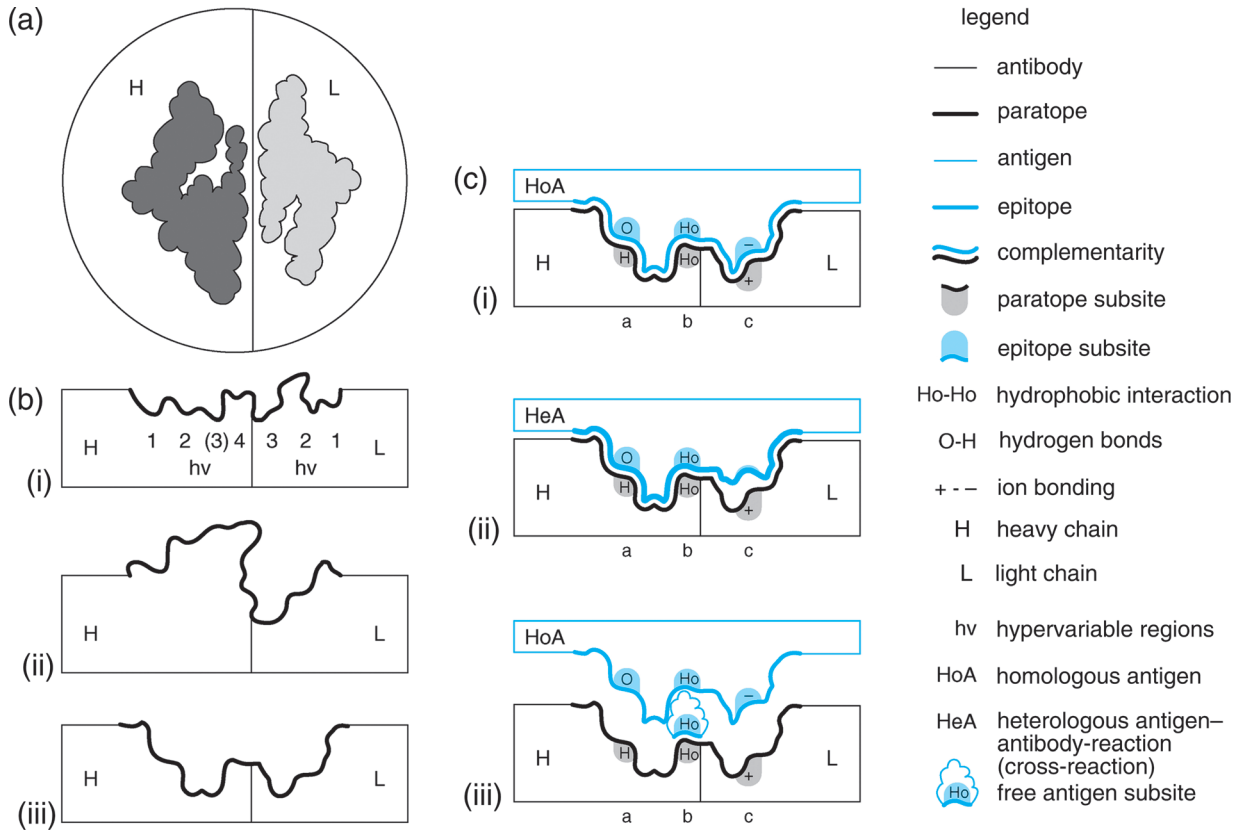


Figure 5.2 Schematic of the combining site, the paratope. (a) Top view: The paratope is formed from the heavy (H-) chain (dark grey) and light (L-) chain (light grey) such that the H-chain possesses the larger part (also functionally) of the paratope as depicted. (b) Side view of three paratopes with different specificities and schematic indications of the hv-regions. The hv regions lie on the surface of the paratope (c) but are distributed three-dimensionally (a, b). The complementarity is also variable, just like the positioning of the various sub-functions within the paratope with the same epitope. (c) Finer details of the epitope–paratope interaction of a homologous antigen–antibody interaction (i) and a heterologous immune reaction with a heterogeneous antigen (ii). Also depicted is the inhibition of a homologue immune reaction by a single dominant sub-epitope (iii).

varied physicochemical functions (Section 5.3). The paratope is variably formed, depending on the specificity of the antibody. The exactness of fit is ensured by the complementarity. For this reason the paratope can display a cavity or else a swelling or another configuration in various parts (Figure 5.2).

How one can visualize an epitope–paratope pairing with its various sub-functions is shown schematically in Figure 5.2c. The stability of interaction is ensured by about six sub-functions (paratope-*sub-sites*) of the antibody (hv-regions, Figures 5.1 and 5.2b) and through the corresponding epitope *sub-sites*, from which three different examples are shown of various bonding mechanisms in Figure 5.2c, showing hydrogen-bonding O—H, hydrophobic bonding Ho—Ho, and ionic or salt bonds.

There are two different modes of immune reactions. In Figure 5.2c an example of the homologous antigen–antibody reaction is depicted, showing a paratope–epitope bonding with stereochemical accurate fitting between epitope and paratope with all three bonding functions complete. Therefore, the bonding force (affinity and avidity, see Section 5.3) is very high.

In contrast, an example of a heterologous antigen–antibody reaction (cross-reaction) is shown in Figure 5.2c(ii), in which a heterologous epitope of a heterologous antigen (HeA) reacts with the same antibody as in Figure 5.2c(i). Because this epitope is only chemically related to, but not identical with, the immunogen through which the antibody used was induced the physicochemical complementarity in Figure 5.2c(ii) is not perfect. In this case the missing salt bonding absent in Figure 5.2c(ii) does not allow a perfect fit and the bonding force is lower than that in Figure 5.2c(i). It is also indicative that the absence of a single sub-bonding function can alter the total conformation of the epitope so much that vicinal *sub-sites* can also become destabilized (c in Figure 5.2c(ii)).

The inhibition of a homologous immune reaction is depicted in Figure 5.2c(iii) as a further function. The epitope is shown with the homologous paratope, as in Figure 5.2c(i). In this case, however, a hv-region of paratope *sub-site* (Ho) is occupied by an isolate epitope-*sub-site* so that the otherwise bonding paratope via a central sub-function is inhibited so much that the homologous reaction between antigen and antibody cannot take place. This inhibition occurs only when a single sub-function of an antigen is so immunodominant that the *other sub-sites* are more or less silenced, which can lead to a complete inhibition of the immune reaction.

This type of immunodominant sub-epitope inhibition is known with haptens, like dinitrophenyl, with protein fragments, or synthesized peptides that can display this type of dominance. A similar suppression of immune reactions is found occasionally also in agglutination-inhibition tests, with extinction phenomena in immunoprecipitation, in competitive radioimmunoassays (Figure 5.17), and in the corresponding ELISA (Figure 5.24).

5.1.6 Handling of Antibodies

The greatest problem concerning handling and storage of antibodies is the loss of antibody activity through structural change. Especially when isolated, pure antibodies tend towards aggregation after freezing and become easily damaged by bacterial growth. Therefore, some suggestions are offered as to how these problems can be avoided. Isolated antibodies can be protected from aggregation and loss by adherence to plastic surfaces by the addition of 1% bovine serum albumin. Bacterial growth is greatly delayed at 4 °C and mostly inhibited by the addition of (strongly toxic) substances like sodium azide (NaN_3, 0.02% final concentration) or Thimerosal (0.01%). Storage in a frozen state after shock freezing (see above) reduces the storage-related antibody loss, so long as it concerns freezable antibodies. Non-freezable antibody preparations are best stored sterile at 4 °C and in NaN_3, with possibly also storage at −20 °C after glycerol addition. For commercial antibodies the appropriate conditions can be found on the instruction leaflet.

Antibodies are most stable in their "natural environment", which means in their original form as antiserum. In this case, freezing and thawing usually is not a problem as long as it is not done too often. To have antibodies of equal activity always available, store the antibodies preferably apportioned and shock frozen at −20 °C or lower in the same way and thaw out each time a new tube for the next usage and store the rest at 4 °C for only a few weeks.

The greatest problem with freezing in small portions and with long storage is the drying out of the samples, which in many cases leads to loss of antibody activity. Drying out is a problem with tubes that are not firmly closed. This can be prevented by the use of a screw-top with a plastic ring and by a Parafilm M® covering with external sealing of the screw thread.

Drying out occurs also with firm stoppering if the space above the sample is too large. The reason for this is the continual temperature variation from the automatic temperature regulation and especially from continual opening and closing of the refrigerator. With each temperature rise the water vapor pressure of the frozen water in the sample increases and a little more liquid evaporates into the space above the sample. During the next cooling phase the air over the sample gives up some humidity again, but to the coldest area of the tube, which is the sample-free area in the top part.

Even when the actual liquid transfer from the sample to the tube wall per cycle seems miniscule, it is visible as the formation of ice crystals on the wall above the sample and as the increased brightness of the drying out sample (Figure 5.3a). For this reason samples stored over long periods should occupy if possible three-quarters of the tube volume. The temperature variations can be minimized when the tubes are kept in a thick box and the samples are best stored in a chest freezer with access from above.

For cryopreservation in the unfrozen state below 0 °C the antibody can be mixed with protective media such as glycerol, in 10–50% final concentration, which do not reduce the major antibody activity and do not prejudice its later use. These media lead, for example, depending on concentration, to lowering of the freezing point, so that the sample is not only protected from ice crystal formation, but also from drying out. Glycerol is therefore mainly employed as cryopreservative for the cold storage of very small amounts of sample (for the aliquoting of quite varied proteins).

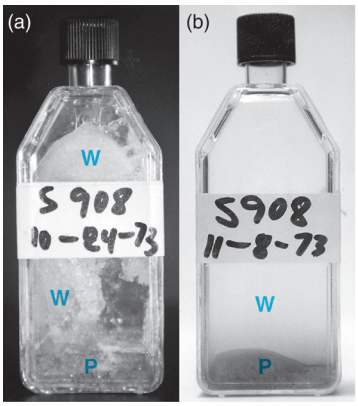


Figure 5.3 Inadequate storage of an antiserum. (a) About 15 ml antiserum was stored in a 40 ml tissue culture flask with a tight screw cap for over 30 years at –20 °C in a refrigerator without safety precautions. The antibody water (W) has separated by reverse sublimation from the antiserum protein (P) in the form of ice crystals. (b) Thawing of another similar flask without shaking shows that the amount of liquid was not reduced, but the protein was separated from the solution, which led to a reduction of antibody activity.

5.2 Antigens

In immunology the term antigen is not defined by a chemical configuration but by the existence of an antibody that with a definite affinity binds a substance. This substance has, thus, one antigenic specificity, defined by the antibody. Antigens are mostly proteins or polysaccharides; they are also, seldomly, glycolipids or other substances.

Antigens that produce an immune response are called complete antigens or immunogens (allergens, tolerogens). The properties of an antigen that are necessary for an immune response are listed in Box 5.2. If a substance is too small to be an immunogen (incomplete antigen, hapten) it can, however, become an immunogen through coupling to an immunogenic vehicle (Box 5.2). Incomplete antigens can come from a great pool of chemically various classes of materials and also from laboratory synthetic substances that do not occur in nature. Haptens can also be much smaller than an epitope.

An antibody reacts only with certain sub-regions of the surface of an antigen, the so-called antigen determinant*s (epitopes)*. Epitopes can be *sequential, conformation-dependent,* and *hidden* or appearing *new* (neo-antigen) after enzymatic cleavage (Figure 5.4). Natural antigens are mostly multivalent, which means that they possess, according to size, several to many antigen determinants.

Epitope An antigen segment bound by a paratope (Figures 5.1 and 5.2).

The epitope consists of about six sub-regions, the so-called *sub-sites*, and each of these can possess a different physicochemical function (Section 5.3). These sub-regions, named paratope sub-sites, correspond to the appropriate individual CDR regions (hv-regions) of the paratope. The individual epitope sub-sites usually have different values for the interaction power of the epitope. The strongest sub-sites are called immunodominant and an unattained interaction power at one of the hyper-variable regions is described as immunosilent. The values of the other epitopes lie within this spectrum and the sum of the sub-sites determine the total interaction strength of an epitope.

An antibody interaction can also lead to a conformation change. This makes it possible to produce antibodies against certain unstable intermediates in the case of enzymatic catalysis. The antibody made in this way can force the corresponding antigen into a conformation that makes the antigen the substrate for a subsequent enzymatic reaction. In this way, the antibodies can possess catalytic properties.

Enzymes as Catalyst, Section 3.4

Catalytic antibodies can be employed for the preparation of certain chemicals by using their enzymatic nature, and also for the establishment of homogenous immunoassays (Section 5.3.3).

Box 5.2 Requirement for the immunogenicity of proteins.

Structures, which determine the immunogenicity of a protein and lead to epitope selection in immunized animals, were identified through artificial antigens (haptenized immunogens) and synthetic peptides of defined amino acid sequence mainly by Michael Sela (before 1970) and later by other research groups.

The following requirements seem to be important:

- Great chemical difference between the immunogens and the immunized individual, which is caused mainly by the evolutionary distance.
- The size of the proteins should be above 5–10 kDa. Amino acids and small peptides under about 30 amino acids are usually not immunogenic. For the induction of antibodies, peptides are coupled covalently to strong immunogenic carriers, for example, on thyroglobulin or the hemocyanin of horseshoe crabs *Limulus polyphemus*. Likewise, one proceeds with chemicals of low molecular weight that are inserted as haptens.
- The immunogen must possess a certain complexity. Polymerized mono-amino acids are in general not immunogenic. A combination of various amino acids is essential. Nevertheless the immunogenic powers of the individual amino acids differ greatly.
- The type of side chains of amino acids influences the immunogenic potential of the epitope. The strongest epitope-determining amino acids are polar (charged amino acids like glutamic acid, aspartic acid, lysine, and arginine) and with large side chains, especially those with rings (tyrosine, phenylalanine, tryptophan).
- The immunogenic regions must be easily accessible on the surface of the molecule for recognition.
- Some well-ordered denaturation and aggregation can increase the immunogenic power.
- The immunogen must be processable for antigen-presentation by antigen-presenting cells like monocytes, macrophages, dendrites and related cells.
- Statistically, the epitopes exhibit a high antigenic index, which is the sum of various numerical components. Immunogenic epitopes are exposed at the surfaces of the molecules (high surface likelihood) with mostly hydrophilic amino acids (low hydrophobicity index). They are flexible (high flexibility index), show no β-structure (low β-structure likelihood) and contain plenty of glycine and proline, which lead to turns (direction reversals along the polypeptide chain).

Epitope mapping of a continuous epitope can be done in various ways. In the past, one started with *limited fragmenting* of the homologous immunogen, first using endoproteases, which normally yield larger protein fragments. The reactive fragment was then further split with various enzymes until the smallest fully reactive fragment was identified. This method is, however, not very precise since it depends on the known predictable cleavage sites of normal enzymes. The testing of synthetic peptides is therefore adopted, which can be produced without the above-mentioned restrictions and which yield more precise results. Nowadays, epitopes are mapped most practicably with a series of *systematic staggered synthetic peptides* from about octapeptides to duodecapeptides, which are commercially produced. Epitope mapping is shown in an example in Figure 5.5. A polypeptide of 26 amino acids should be the immunogen of a monoclonal antibody whose epitope is sought by means of hendecapeptides that have three amino acids continually replaced to cover the complete polypeptide. The peptides 7–17 and 10–20 react fully, which shows that the epitope must lie within positions 10–17. A shortening of this peptide from the N- and C-terminal end shows

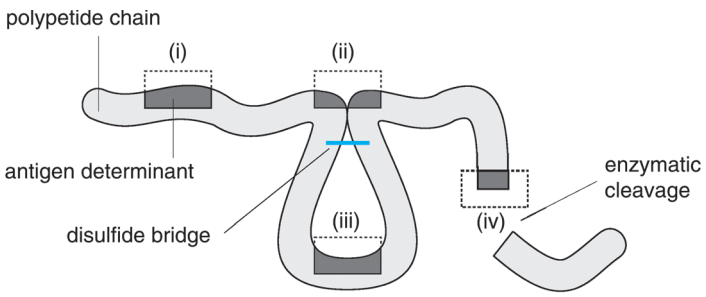


Figure 5.4 Antigen determinants, which are about 5–8 amino acids in size, in various ways form an epitope: (i) sequential or continuous epitopes; (ii) conformational or discontinuous epitopes – they can be destroyed through alteration of the conformation as a result of denaturation or by disulfide cleavage; (iii) hidden antigen determinants, which can, for example, become exposed by denaturing or by disulfide cleavage; (iv) by proteolysis to newly made antigenic determinants (neo-antigens).

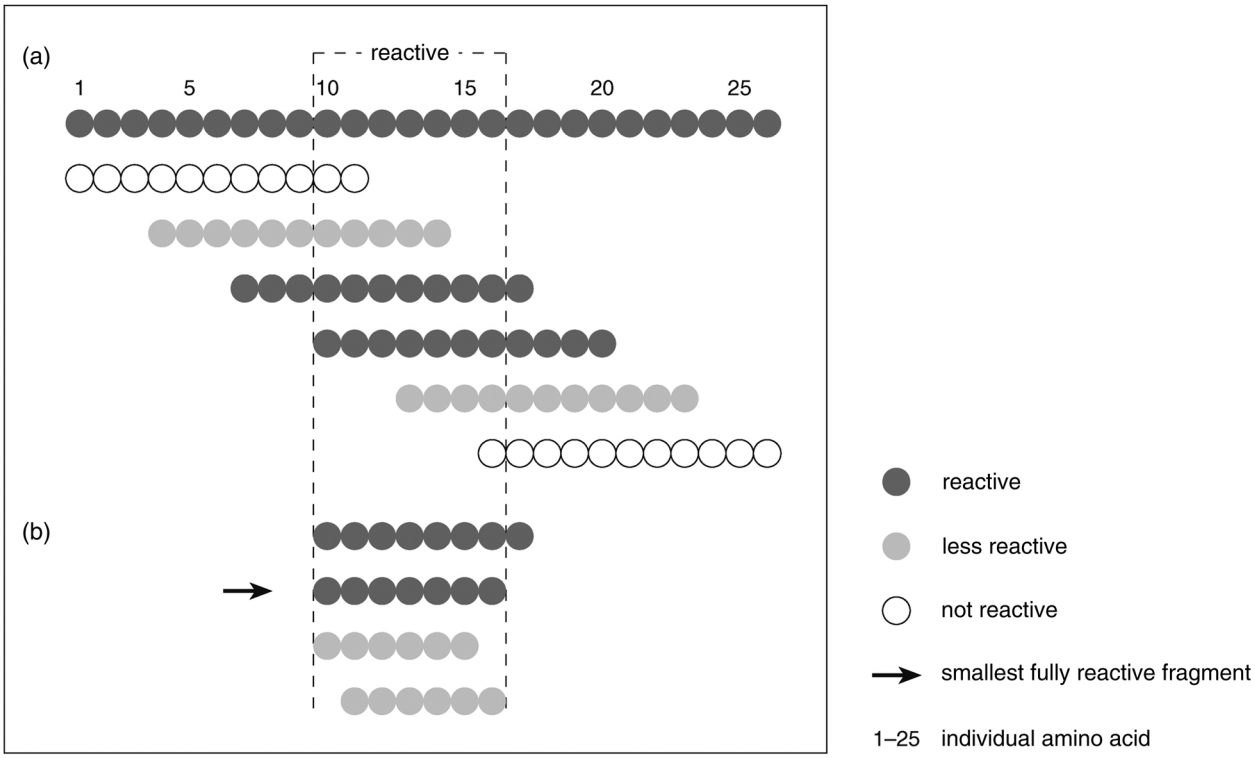


Figure 5.5 Epitope mapping of a monoclonal antibody: (a) exploratory mapping; (b) fine tuning.

only with peptide 10–16 the full immunoreactivity with the monoclonal antibody; therefore, this represents the sought epitope.

5.3 Antigen–Antibody Reaction

The forces effective in the antigen–antibody reaction are the same as, for example, those effective for the ligand receptor, the enzyme–substrate interaction, and for safeguarding the conformation of proteins; these are stereochemical complementarity, hydrogen bonds and ion bonding, van der Waals forces, and hydrophobic interactions (Figure 5.2c). The specificity of an antibody is dependent on the bonding strength (avidity) to its antigen. The avidity is composed of several components:

- The affinity, which is defined by the monovalent interaction strength between epitope and paratope (over all *sub-sites*) and is measured by the determination of the equilibrium constant.
- The multivalency of antibodies.
- The often overlooked Fc-association between different IgG-molecules after antigen interaction (Figure 5.6b(ii)).

While the equilibrium constants exist as chemical units, and the antigens (a single conformation) as well as the antibody (monoclonal if possible, Figure 5.7)) can be measured exactly by equilibrium dialysis (and other methods), the avidity remains a more functional concept that is not so easily defined. The avidity is nevertheless attainable via the speed of the antigen–antibody reaction in comparison with monovalent Fab-fragments and divalent $F(ab')_2$-fragments (Section 5.3.3.).

The affinity rests on the stereochemical fitting accuracy (complementarity) of epitope and paratope (Figure 5.2c) and the above-mentioned interaction functions. For the accuracy of a fit, the flexibility and mobility of epitope and paratope play an important role for the final fit. The affinity can change strongly. Homologous antigens (= used immunogens) possess the strongest affinity. Cross-reacting antigens (heterologous antigens) that are not identical to the immunogens but only chemically related usually have less affinity (Figure 5.2c).

In rare cases when the heterologous antigen binds more strongly than the homologue the antigen is classified as heteroclite (abnormal or irregular). This can be due to an antibody-dependent conformation change of the antigen resulting in exposure of previously hidden antigen determinants (Figure 5.4). Since interactions of low affinity exist owing to epitope and paratope partial complementarity and flexibility one must set stringent requirements to reduce or even exclude such interactions of low affinity, that is, the paratope-mediated "unspecific" interaction.

Historically, the antigen–antibody reaction has been visualized using various methods. Depending on the type of the antigen the visible result has been described as agglutination, precipitation, immune binding, and complement fixation. While the first two can be seen on inspection, immune interaction and complement fixation are first made visible by means of amplification systems. Each of the four immunological types of verification follow their own rules and possess their own field of application. For this reason the special characteristics of the four types of immune reactions and the subsequent reactions will be briefly presented and their field of use explained in the following sections.

5.3.1 Immunoagglutination

When antigen-bearing microscopic, suspended particles, like bacteria or blood cells, or antigen-bearing particles are mixed with a corresponding immune serum, the corpuscles clump together. The result is instability of the suspension and, finally, visible sedimentation. This form of antigen–antibody reaction is named agglutination.

Direct agglutination (agglutination via the primary antibody) is based on the crosslinking of antigens at the surface of corpuscles by antibodies, such that a single antibody molecule with its identical paratopes must bridge over to reach the same antigen determinants on different cells. Clearly, for this bridging of two and more particles the IgM antibody is most suitable owing to its size and decavalency, whereas the monomeric bivalent IgG antibody is not so suitable. If the IgG antibody does not agglutinate, the use of an additional antibody that is directed against IgG (secondary antibody) can lead to agglutination. This form of agglutination is known as indirect agglutination (agglutination via a secondary antibody). It plays a role in the clinic for the diagnosis of pathogenic antibodies against cell surface antigens.

An excessive antibody overloading leads, however, to inhibition of agglutination since statistically each epitope now interacts with a single antibody and so makes crosslinking impossible. To prevent this so-called prozone effect, which can appear in all immune assays, an immune reaction with an unknown antibody (or with an unknown serum) must first be optimized in an appropriate dilution series to identify the "specific window" at the "zone of equivalence" (see below).

Application Direct agglutination is applied for identification of cell surface antigens, such as blood group determination by means of hemagglutination. Any antigens can also be attached covalently to erythrocytes and, thereby, "passive immunoagglutination" (direct or indirect) can be carried out *in vitro*. Additionally, by means of antibodies attached to erythrocytes or latex particles – this variant is known as "reverse immunoagglutination" – dissolved antigens can within minutes be qualitatively (and through comparative dilution series, semi-quantitatively) identified. In addition, agglutination inhibition assays are especially useful with monovalent antigens, antigen fragments, individual immunodominant epitope-*sub-sites*, or with mutation-altered epitopes (Figure 5.2c(iii)).

Hemagglutination is an exquisitely sensitive test for antibody titer determination. If, for example, one lets diluted erythrocytes in a small plastic funnel (special microtiter plates) run down the circular steep plane of the funnel wall to the bottom of the funnel, an interaction of the antibody to the erythrocytes (and in particular the crosslinking) is visible even with a few antibodies due to the inhibition of their flow.

Although immunoagglutination can only deliver semi-quantitative data, because of its high sensitivity and great versatility, but mainly because of its relatively simple application and low cost, it is of great value in many investigations. With the development of other extremely sensitive and exact binding tests (Section 5.3.3) the importance of immune agglutination has greatly decreased, except for its application in medicine.

Titer An antibody titer is the highest dilution step at which an antibody function is still measurable. The titer is used for a practical comparison of antibody effects. The titer depends on the antibody concentration, its affinity, and avidity (Section 5.1.3), but mainly also on the type and sensitivity of the detection system used.

5.3.2 Immunoprecipitation

A mixing of soluble antigen with specific antibodies leads to an antigen–antibody reaction that, by loss of solubility of the antigen–antibody complex, leads to cloudiness followed by sedimentation. This process, which resembles agglutination except that here both partners are soluble, is traditionally known as immunoprecipitation. A precipitation can only occur if at least three (in some cases four) antigenic determinants are available on an antigen.

Co-immunoprecipitation, Section 16.4

As with agglutination, there can also be a marked prozone effect in the case of precipitation. The quantitative study of this effect was published by Larsen in 1922 and later by Michael Heidelberger and Forrest E. Kendall in 1937 (Figure 5.6). For a series of test tubes containing the same amount of antibody, from left to right an increasing amount of antigen is added and at some time following successful antigen–antibody reaction the precipitate is measured. It is then seen that a defined relationship of antigen to antibody gives the maximum yield of immunoprecipitate (Figure 5.6a).

If the amount of soluble antigen or active antibody is then investigated in the supernatant, the tube with maximal precipitation contains neither antigen nor antibody, which means that antibody and antigen are *in toto* crosslinked, while the tubes to the left of this contain active

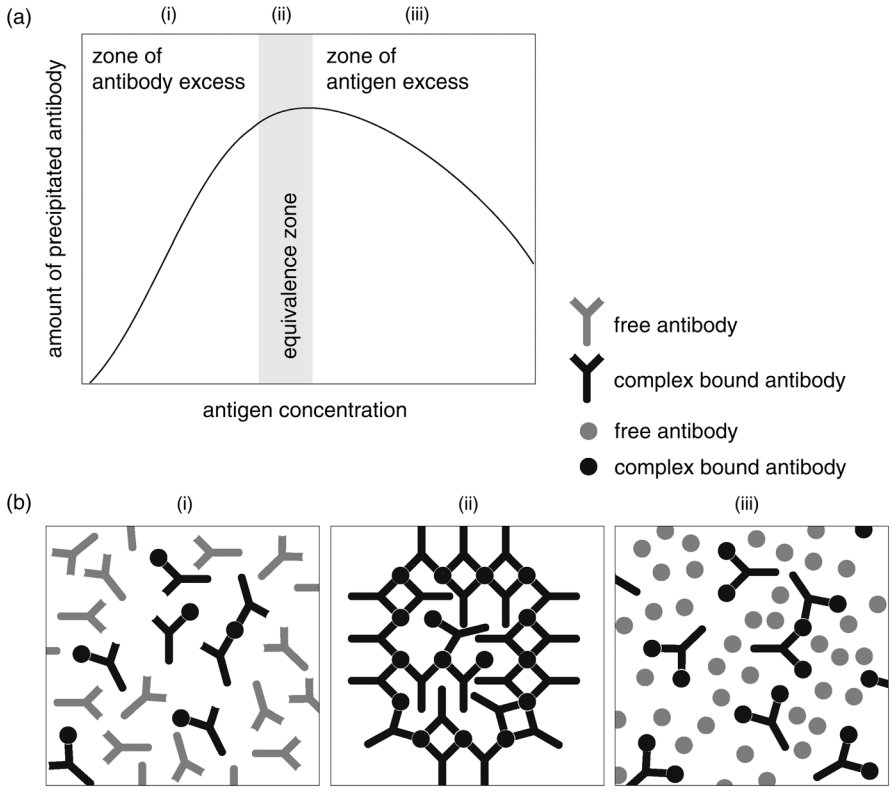


Figure 5.6 Quantitative immunoprecipitation (Heidelberger curve). (a) Determination of precipitation optimum through measurement of the amount of precipitated antibody as well as that of antigen and antibody activity in supernatant after the end of an immune precipitation run. (b) Visualization of individual partners of the antigen–antibody reaction with different antigen/antibody relationships. Precipitation at the point of equivalence in (ii) is conditional on the generation of a spatial crystal lattice of undetermined size, which is insoluble; (i) soluble immune complexes in antibody, (iii) in antigen excess.

antibody (zone of antibody excess) and those to the right contain soluble antigen (zone of antigen excess). The zone with the maximal precipitation is called the zone of equivalence, which means the equivalence of epitope and paratope. The existence of an equivalence identifies a stoichiometric relationship between epitope and paratope, a premise on which the use of quantitative immunoassays is based (see below). The maximal precipitation with epitope and paratope equivalence is of course different for every antigen–antibody system ("serological system" or "antigenic system") and depends on the valency of the antigen as well as the number of available paratopes in an antiserum. It is, therefore, clear that the *exact* point of equivalence cannot be calculated from the molarities of the reaction partners.

Antigenic system This refers to the description of the antigen–antibody system in question. An antigenic system consists of an antigen and all reacting antibodies within the given system. A synonym is a "serological system" (Jacques Oudin 1946).

Figure 5.6b illustrates how the immunoprecipitates are formed in the three zones ((i), (ii), (iii)). Both in antibody excess and in antigen excess there are small soluble immune complexes and little precipitation, whereas at the point of equivalence after the interaction of all antigens with all antibodies the formation of maximal immune complexes in the form of a spatial network is completed. The crosslinked reaction partners fall out of solution in the form of clearly visible precipitates.

The term "antigenic system" is illustrated in Figure 5.7. The figure shows minimal systems of antigen–antibody reactions at the single epitope–paratope level.

Figure 5.7a shows a precipitating antigen system with three epitopes (1,2,3) and three corresponding antibodies (→1, →2, →3), which can be polyclonal or monoclonal. A single antibody or peptide antibody (→1) directed against an epitope (1) cannot precipitate with the antigen depicted in (a) (Figure 5.7b), but only in the presence of three identical (repetitive) epitopes on an antigen (Figure 5.7c). Because (with the exception of some structural proteins carrying repetitive antigenic determinants) the naturally soluble antigens as a rule carry no identical or repetitive epitopes, the corresponding monoclonal or peptide antibodies cannot precipitate with these natural antigens.

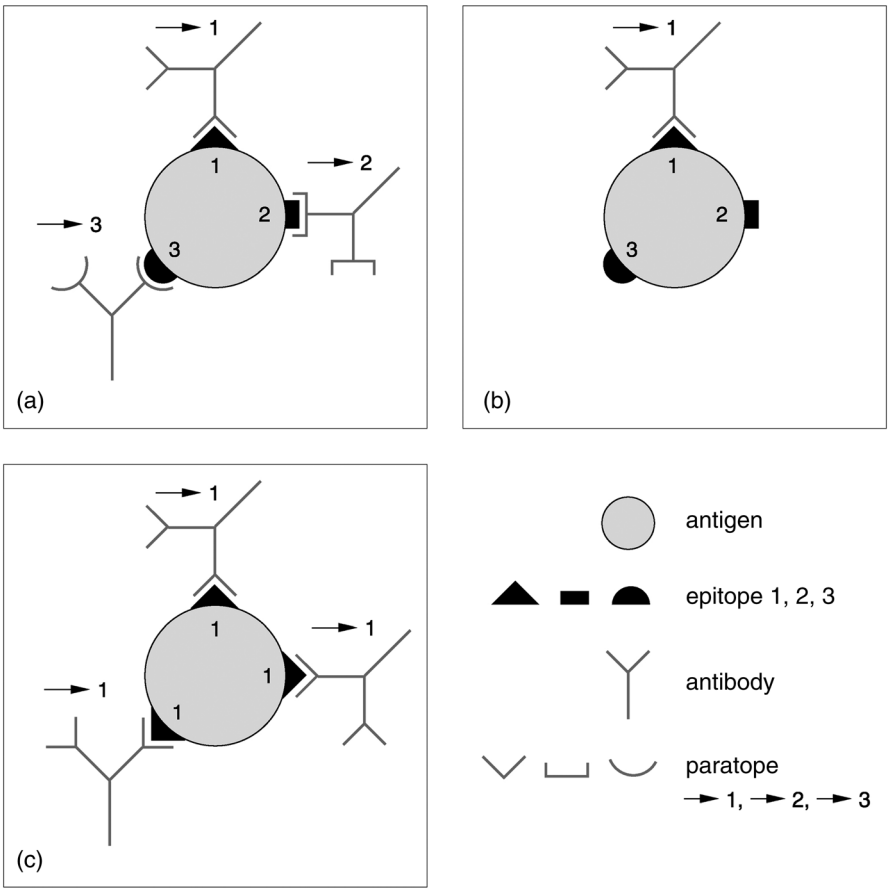


Figure 5.7 Minimal models of antigen systems. (a) Precipitating system, consisting of an antigen with three antigen determinants (epitopes 1–3) and three matching antibodies (→1 to →3). The antibodies can come from an antiserum or represent three monoclonal antibodies. (b) Non-precipitating system, which consists of the same antigen as in (a) and a monoclonal antibody against one of the epitopes (1, →1). (c) Precipitating system, with three identical (repetitive) antigen determinants and monoclonal antibodies of identical specificity.

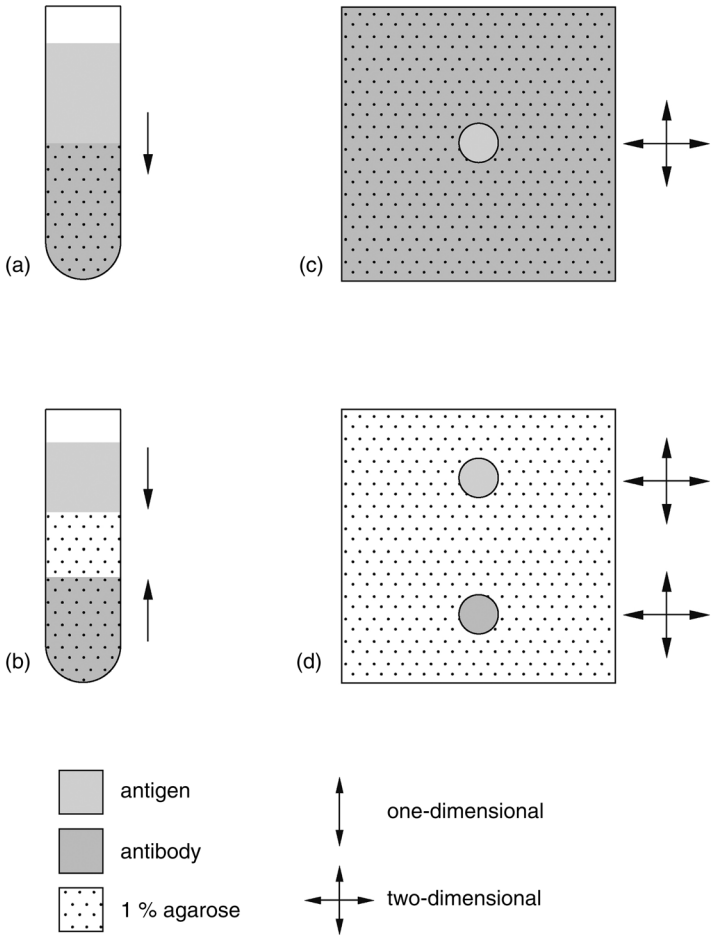


Figure 5.8 Classification of immunoprecipitating systems. The four basic arrangements are shown immediately after application of the reactants, at the start of diffusion. (a) One-dimensional simple immunodiffusion according to Oudin. (b) One-dimensional double diffusion according to Oakley and Fulthorpe. (c) Two-dimensional simple (radial) diffusion according to Mancini. (d) Two-dimensional double diffusion according to Ouchterlony.

The history of analytical immunoprecipitation began at the end of the nineteenth century with a glass tube in which an antiserum was overlaid with an antigen solution. After some time the diffusion of both reaction partners into each other gave in the vicinity of the boundary layer a white disc that corresponds to a specific immunoprecipitation. This test, the so-called ring test, was the starting point of all immunoprecipitation methods, of which there are basically four different configurations, which are illustrated in Figure 5.8 and in the following with important examples shown in detail. To succeed, a precipitation requires antigens that are soluble in physiological buffers. To dissolve insoluble antigens (lipoproteins, membrane proteins, protein fragments) in physiological buffers under certain conditions that do not materially affect the activity of the antibody, salts in high molar concentrations (e.g., guanidine·HCl), urea, and/or detergents can be used. Since artefacts can occur in the immunoprecipitation with this type of solution, corresponding checks are necessary. Immunodiffusion with the help of these denaturing solvents can only succeed if the antigens are in high concentration and the volume of solvent is small. Since the solubilizing molecules are very small and therefore diffuse rapidly they are sufficiently diluted by diffusion before antigen and antibody meet each other.

One-Dimensional Simple Immunodiffusion of Oudin In 1956, Jacques Oudin brought important knowledge to the understanding of immunoprecipitation. Some of his now historic experiments are worth mentioning for instructional purposes. The investigation of precipitation in solution (ring test, see above) is difficult because of the instability of the fluids, the higher density of the resulting precipitate, and temperature-induced convection. Oudin, therefore, used

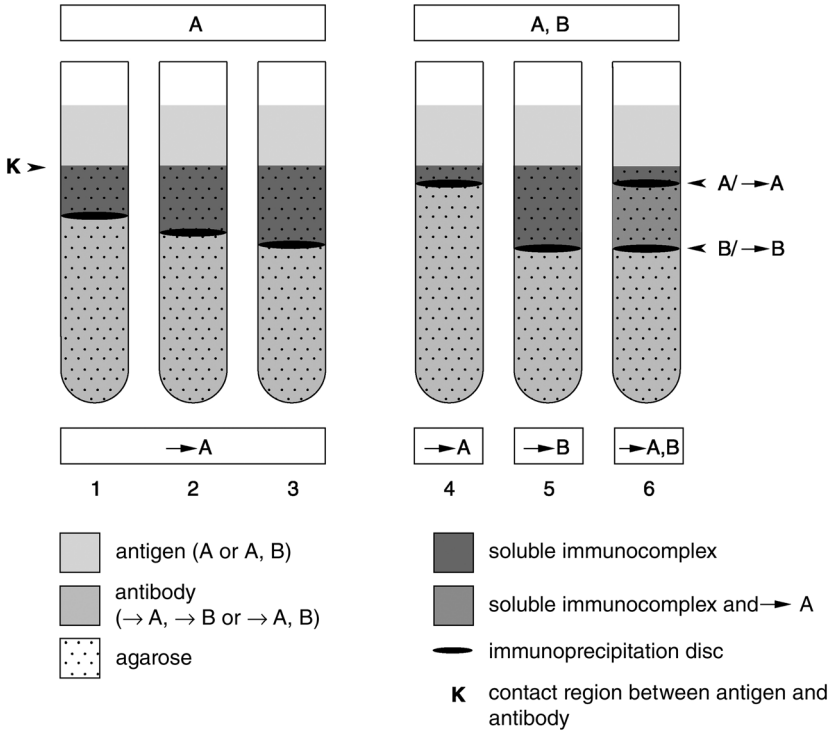


Figure 5.9 One-dimensional diffusion according to Oudin. A, antigen A; B, antigen B; →A, anti-A; →B, anti-B; A/→A, antigenic system (see Figure 5.7).

as reaction partner mostly antiserum, cast into wide-meshed gels, and thus allowed free diffusion and gained stabilized immune precipitates (Figure 5.8a) of larger reaction partners greater than 1000 kDa; on the other hand, they fix and stabilize the resulting immunoprecipitate in the gel meshes. Today, one uses 0.5–1.5% agarose in physiological buffer as stabilizing gel. In Figure 5.8a, the antibody trapped in gel is overlaid with liquid antigen and each arrangement is depicted in Figure 5.9 to show the important insights revealed concerning the principles of immunoprecipitation.

If the same amount of gel-stabilized antibody →A is overlaid with a solution of increasing concentration of antigen A (Figure 5.9, tubes 1–3), then with increasing antigen concentration the precipitation line forms at a greater separation from the contact region K between A and →A. This greater separation with higher concentration can be explained as follows: The antigen concentration sinks along the diffusion path in the equally distributed gel-stabilized antibody the antigen. The initial excess antigen immunocomplex becomes ever more saturated with antigen with further diffusion, remains partly held in the agarose, is partly resolubilized by the following antigen, diffuses further, and finally precipitates out at the equivalence point. If more antigen is available, it needs stronger dilution, which is achieved with a larger diffusion distance. Because of the stoichiometry of epitope and paratope this distance is proportional to the antigen concentration. In this way Oudin achieved the first exact concentration determination of individual antigens in a complex protein mixture by using a monospecific antiserum through the quantitative immunoprecipitation.

Antisera →A and →B lead to two precipitation lines that are independent of each other since they behave the same way as in separate tubes (Figure 5.9, tubes 4, 5).

The variation of concentration of *one* antigen leads only to displacement of the corresponding precipitation line, as shown already in Figure 5.9 (tubes 1–3), and not to displacement of the other. From this comes the important insight that antigen systems in the same tube precipitate independently from each other (Figure 5.9, tube 6). Because the precipitation lines in Oudin's system wander within the antibody compartment and are not stable, the simple diffusion is replaced by an apposed ("double") diffusion (Figure 5.8b). In this system of Oakley and Fulthorpe (1953) the concentration of *both* partners decreases

during diffusion in a protein-free gel segment. Because of this, the precipitation optimum is relatively stable, resulting in sharper precipitation lines at best resolution. This principle was also realized by Ouchterlony in two-dimensional immunodiffusion.

Two-Dimensional Immunodiffusion of Ouchterlony With the technique specified by Örjan Ouchterlony in 1948, antigen and antibody each fill a punched hole, (*well*) in an empty agarose plate (e.g., 1% agarose, physiological buffer) and are diffused against each other (Figure 5.8d). During diffusion, both partners form a concentration gradient, making a sharp precipitation line where they encounter epitope and paratope equivalence. This (still one-dimensional) double diffusion follows Oudin's rule – nevertheless the result can be read with much less effort. The addition of a third well enables diffusion in the second dimension and, with this, an antigen comparison as illustrated (Figure 5.10(i)–(iii)).

The Ouchterlony technique is the simplest technique of high precision with which especially proteins and antibodies can be reliably identified and compared in a short time with a minimum of effort since antigens and antibodies can react at a wide range of concentrations while finding their optimal precipitation at the epitope–paratope-equivalent by the steep diffusion gradients they are themselves producing.

Figure 5.11 shows an example of an original immunodiffusion with human serum, which represents a mixture of many proteins with five predominant serum proteins identified by comparative precipitations with isolated proteins as albumin (left-hand side) and IgG (right-hand side).

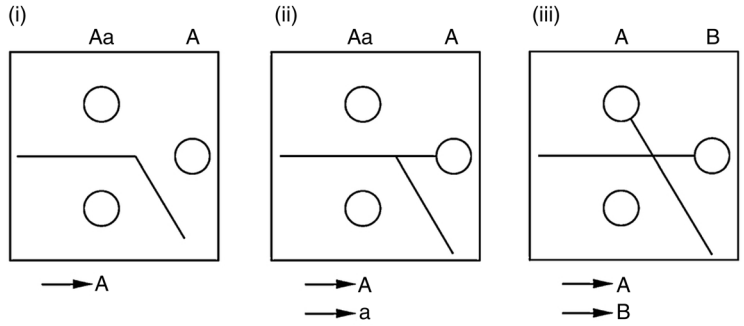


Figure 5.10 Comparative two-dimensional double diffusion according to Ouchterlony (Figure 5.8d). Agarose gel with punched holes is filled with the reagents indicated. The two holes contain the different antigens and one hole contains the different antibodies. There are three basic precipitation patterns to be distinguished. (i) Precipitation of the line of identity: The precipitation lines between the antigen A and Aa (Aa is larger by the polypeptide a, see (ii)) fuse in a continuum, meaning that the epitope–paratope equivalent is the same in the two antigens A and Aa and the segment of A and Aa is shown to be identical by the anti-A antiserum. (ii) Precipitation with a line of partial identity: When the same pattern and the same antigens are applied as in (i) but the antibody has the added specificity of anti-a, the A specificity in both (A and Aa) precipitates with a line of identity as in (i) using only the anti-A specificity. The additional anti-a specificity, however, precipitates with a spur of Aa over A (technical term) meaning that Aa is larger than A by the addition of the a specificity. This spur runs into A and by this property shows its *non-identity* with A while the anti-A specificity was absorbed already by being precipitated in the A–anti-A line, letting the anti-a specificity pass through this line and meeting the a-specificity in the spur, which represents the precipitation line Aa and anti-a. (iii) Precipitation line of no identity: The same pattern as in (i) and (ii) but with two different antigens A and B and the two different respective antibodies anti-A and anti-B. The precipitation lines run across each other and run into the opposite antigen hole (as in the case of Aa-a-spur in (ii)). They therefore behave totally independently and show that different "antigenic systems" (Oudin) do not influence each other.

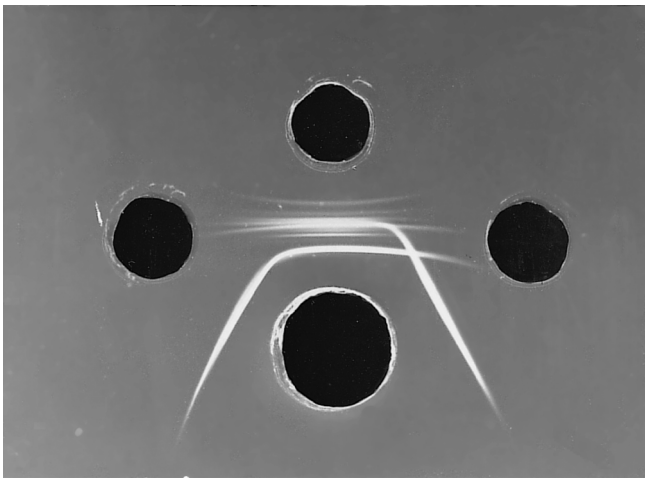


Figure 5.11 Example of an Ouchterlony immunodiffusion: agarose plate (1.5% agarose in 0.1 M tris buffer, pH 7.4) with wells of 1.0 and 1.5 mm diameter. The wells contain, top: NHS (normal human serum), diluted 1 : 10; left: human albumin, 5 mg ml^{-1}; right: human IgG (isolated from serum), 3 mg ml^{-1}; bottom: →NHS, produced in a goat.

Immune- and Crossover-electrophoresis, Electrodiffusion If a mixture containing many antigens is subjected to immunodiffusion, then the precipitation pattern can be difficult to analyze owing to the abundance of different precipitation lines (Figure 5.12a). For this reason the proteins are first separated in an electrical field (Figure 5.12b) and afterwards subjected to immunodiffusion. In the process the well for applying antibody is widened into a channel as shown in Figure 5.12c. This immunoelectrophoresis can separate, depending on antiserum and arrangement, about 15–30 serum proteins. It serves for orientation according to the abundance of certain proteins and the amount as well as the orientation according to the

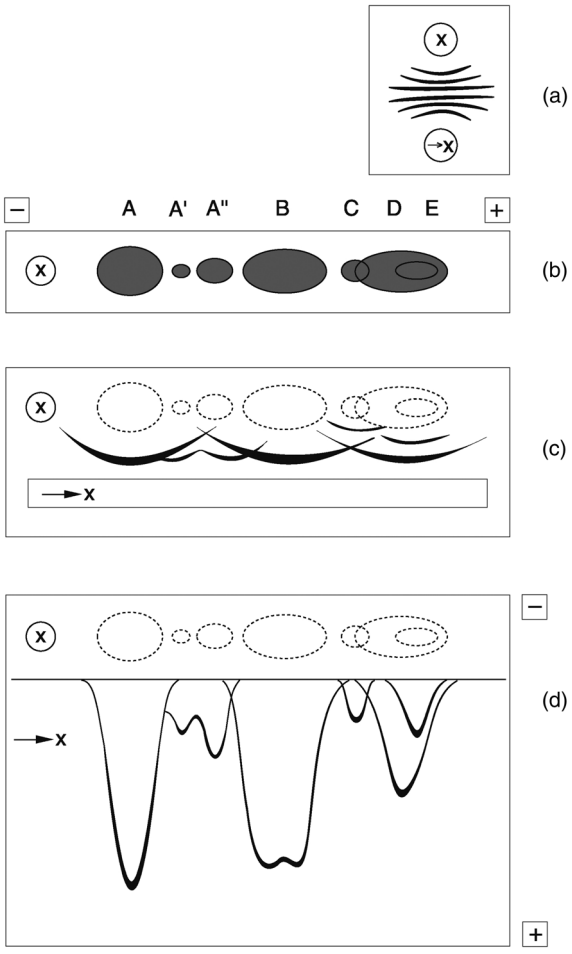


Figure 5.12 Immuno- and crossover-electrophoresis: (a) immunodiffusion with many antigens X; X: Complex antigen mixture; →X: polyclonal antiserum against X. There are many precipitation lines that are very difficult to distinguish. (b) Separation of the protein mixture in X in an electrical field. The proteins are stained in the gel. One can recognize 5–6 protein spots, which represent proteins with different electrical charge. The antigen content of these proteins is analyzed in (c) and (d). (c) Immunoelectrophoresis: Fractionation of the antigen mixture X as in (b) and subsequent immunodiffusion of the separated proteins using →X. The dotted ovals correspond to the position of proteins in (b) (stained with Amido Schwarz 10B), immediately after the start of immunodiffusion. After sufficient diffusion seven different precipitation arcs and their relationship to each other can be recognized: non-identity (B/D, C/D, E/D, B/A″), identity (A′/A″), partial identity (A/A′, whereby A spurs over A′). (d) Crossover-immunoelectrophoresis after Clark and Freedman. Proteins fractionated in the first dimension are electrophoresed from minus (–) to plus (+) in the second dimension into a gel containing →X. The antigenic systems appear as precipitation peaks, the area of which can be used for (semi-quantitative) determination of the concentration of individual components. The antigen comparison (identical, not identical, and partial identity) is analogous to that of immunoelectrophoresis but easier to recognize.

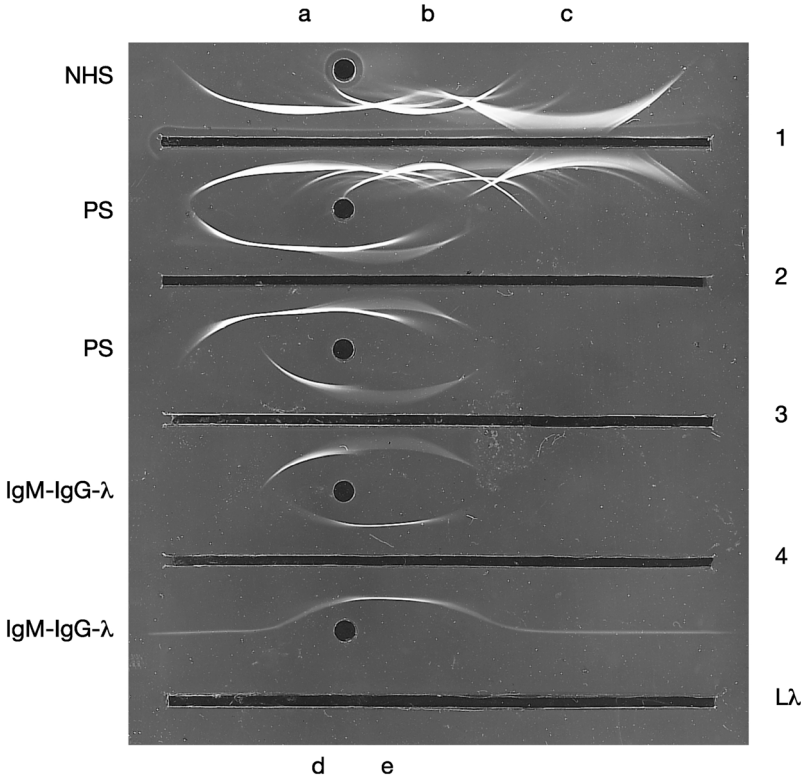


Figure 5.13 Example of immunoelectrophoresis. The separation of a patient's serum containing a monoclonal IgMλ-IgGλ double-myeloma (a protein complex of two monoclonal immunoglobulins) in an electric field. The anode is on the right-hand side. The initial filling point for application of the patient's serum or the isolated complex is the well. *Antigens:* NHS: normal human serum; PS: patient serum containing an IgM-IgG protein complex, isolated by preparative block electrophoresis; Lλ: isolated immunoglobulin-λ-light chain. *Antibodies:* 1, Anti-NHS (normal human serum); 2, Anti-IgG, γ-specific; 3, Anti-IgM, μ-specific; 4, Anti-λ-immunoglobulin-light chain.

charge alterations or the identification of monoclonal immunoglobulins. One can also by means of monospecific antisera identify single protein components in complex protein mixtures (plasma, cytosol, bacterial lysates) and investigate their activation or fragmentation as far as disintegration under defined experimental or clinical conditions. An example of immunoelectrophoresis is shown in Figure 5.13. A patient serum with a monoclonal IgMλ-IgGλ →IgMλ-immunocomplex (monoclonal rheumatoid factor). The complexed IgG is also monoclonal (double myeloma protein).

Electrophoresis, Chapter 11

The resulting precipitation curves in Figure 5.13 correspond to IgG (a), which shows the usual marked charge heterogeneity in normal human serum and in patient serum, transferrin (b), and albumin (c) with a single precipitation curve each, monoclonal IgM (d), and IgM complexed with monoclonal IgG (e). The monoclonality of IgM and IgG is recognizable in the reduced heterogeneity and at stronger concentration. The antigen comparison shows lines of identity with all loading variants of IgG, which are made specifically visible with →IgG (applied in channel 2). In addition, the monoclonal IgM is identical with the IgG complexed IgM, if it precipitates with →IgM. All other lines are not identical, as can be seen in the behavior of a, b and c, apart from the partial identity of the IgM-IgG-complex with the polyclonal IgG, when it also precipitates with →IgG. Here the polyclonal IgG spurs over the monoclonal, possibly because of the greater heterogeneity of the former.

The isolated immune complex shows the same behavior as that in serum. The λ-light-chain put into channel 5 precipitates as a straight line with the →λ-light chain serum put into channel 4 but nevertheless shows a bulge in the direction of the antiserum channel. Actually, the bulge of the →λ/λ line corresponds exactly to the position of the IgM-IgG-complex, whose light chain is thus determined. This precipitation pattern is an example of comparison immunoelectrophoresis, of which there are quite a number of variants.

To accelerate the immunodiffusion one can let the antigens and antibodies move towards each other in the electrical field, provided that both have different charges. This so-called electrodiffusion (see below and Figure 5.16) has the further advantage of increased sensitivity in that the whole antigen runs in one direction towards the antibody.

This is in contradiction to the two-dimensional immunodiffusion of Ouchterlony (Figures 5.10 and 5.11) where the antigen and antibodies diffuse out in two dimensions,

whereby only those parts of antigen and antibody that happen to face each other serve the precipitation.

Electrophoretic fractionated protein mixtures (as in Figure 5.12b) can, by means of electrodiffusion, be electrophoresed in an antibody-containing two-dimensional gel. This variant is named two-dimensional immunoelectrophoresis or crossover electrophoresis according to Clark and Freedman as sketched in Figure 5.12d. Depending on the number of antigenic systems in a complex mixture precipitation peaks appear that, according to the quality of the antisera, allow the quantification of the area under the individual peaks, for example, in sera, cell extracts, and from membrane proteins, the latter also in non-ionic detergents.

With the help of the precipitation lines of identity, non-identity, and partial-identity, the individual proteins are assigned according to the criteria mentioned in Figures 5.10 and 5.12. In addition, the last two techniques can only be successful if the antigen and antibody possess different charges, so that they run towards each other in the electric field. If this is not the case, one can give the antibodies negative charges by means of carbamoylation.

Chemical Modification of Proteins, Chapter 6

Immune Fixation Immune fixation is also a precise immunoprecipitation method for the identification of proteins in complex mixtures that are fractionated in an electrical field (electrophoresis, isoelectric-focusing). In this method the fractionated proteins diffuse out of the separation gel (e.g., agarose) into a laid-over abutting cellulose acetate membrane that is saturated with specific precipitating antibodies. Thus, the diffused antigen precipitates into the meshes of the membrane, which, in contrast to nitrocellulose or immobilon, scarcely adsorbs proteins. This membrane is finally freed from soluble proteins by washing in a physiological buffer and the immune complexes imprisoned in the meshwork are stained with Amido Schwarz. For optimization of immune fixation, the antigen amount, antibody dilution, exposure time, and the following antigen diffusion must be so varied that the precipitation optimum lies in the antibody-containing membrane. Immune fixation is used, for example, to identify various allotypic plasma and serum proteins in different individuals. Allotype determinations are important in anthropology, human genetics, forensic medicine, and in paternity identification. Immune fixation is also used to discover monoclonal antibodies in serum and urine from patients with multiple myeloma. An instance of allotype determination using immune fixation is shown using the example of vitamin-D-binding serum protein in Figure 5.14.

Isoelectric Focusing, Section 11.3.12

Radial Immune Diffusion and Rocket Immunoelectrophoresis Simple two-dimensional diffusion of soluble antigens in an agarose-stabilized antibody is the radial immune diffusion established by Giulana Mancini, 1965. The principle is shown in Figure 5.8c, and in more detail in Figure 5.15. The result shows the formation of precipitate rings at the point of epi and paratope equivalence after about one to two days diffusion time at room temperature. The

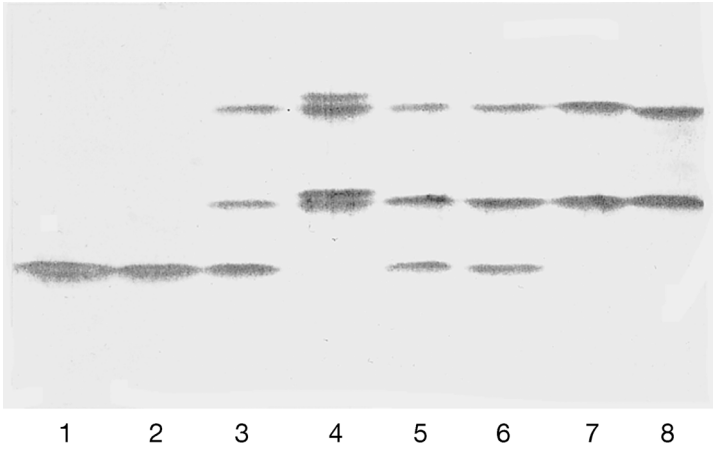


Figure 5.14 Immune fixation as shown by the example of allotype determination of vitamin-D-binding serum proteins (Gc-protein). Whole serum from various individuals* was fractionated in isoelectric focusing and with the help of immune precipitation fixed in an antibody-containing cellulose acetate membrane in the form of an immune complex. The anode is above. Known and visible allotypes of the Gc-protein: 1F–1F (not shown); 1S–1S: 7, 8; 1F–1S: 4; 1F–2 (not shown); 1S–2: 3, 5, 6; 2–2: 1, 2; (F = fast, S = slow).
*Students of the zero-anthropological practical course at the Ludwig Maximilian University in Munich, Summer Semester 1995.

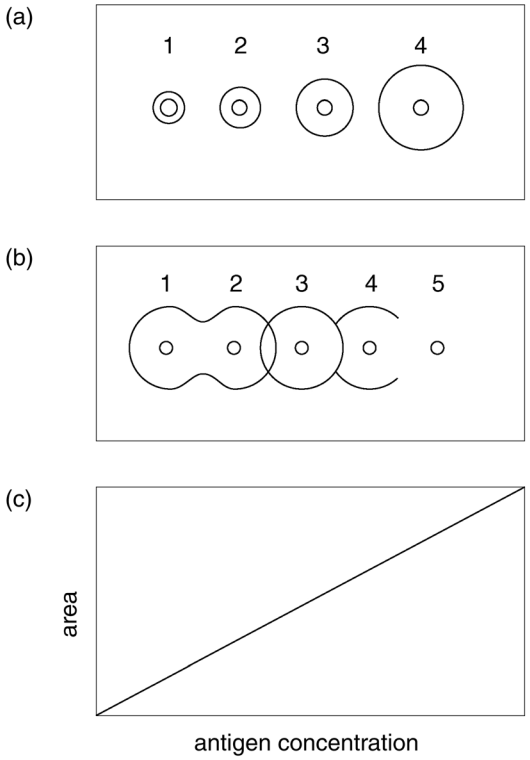


Figure 5.15 Radial immunodiffusion after Mancini. This is an immunoprecipitation method for the quantification of soluble antigens, even in complex protein mixtures. (a) The gel contains a specific antiserum against a single protein. After application of standard antigens of defined increasing concentration in wells 1–4 one obtains stable precipitation rings whose area plotted against concentration yields a linear function after about two days of diffusion (c). (b) Precipitation rings of different proteins, obtained with an antiserum against these antigens, show antigenic behavior of proteins towards each other: Identity between antigen 1 and 2 (short: 1 and 2), non-identity 2 and 3, partial identity 3 and 4, whereby 3 shows a spur over 4, and the extinction phenomenon in 4 through antigen 5 (compare with Figure 5.10).

area enclosed by the rings is proportional to the amount of applied antigen (Figure 5.15a). After calibration with the help of a series of samples with known protein content and the establishment of a standard curve, individual proteins in complex solutions can be quantified with high precision in a simple manner, based on the stoichiometry of epi- and paratope (Figure 5.15c). Various antigen systems, recognizable by the visible appearance of several concentric rings, can be precisely identified by using isolated proteins through rings of identity, non-identity, partial identity (Figure 5.15b), or the extinction phenomenon by using immunodominant single epitopes or other chemically related protein fragments (Figure 5.15b, blocking of 4 by Ho; Figure 5.2c(iii)).

The time required for immune diffusion can be reduced to several hours if antigens move through an antibody-containing gel in an electrical field. With this method appear rocket-like precipitation patterns (Figure 5.16). The possibility of using this electrodiffusion method relies on different isoelectric points for antigen and antibody, so that they can run towards each other in the electrical field. If this is not the case, the antibody can be given a lower isoelectric point by carbamoylation (see above). The peak height of precipitates is proportional to the antigen concentration (Figure 5.16). Rocket immunoelectrophoresis is well suited for large-scale investigations of a series of serums and antibodies.

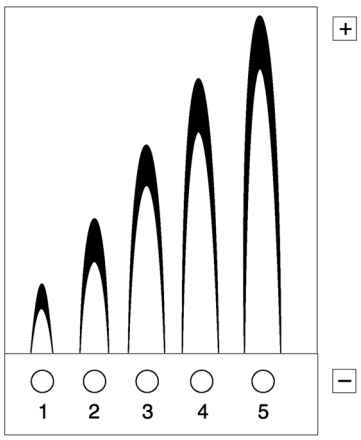


Figure 5.16 Rocket immunoelectrophoresis after Laurell, an immunoprecipitation method for quantitative determination of soluble antigens, even in complex antigen mixtures. Standard concentrations of a protein (1–5) were electrophoresed in a gel-containing antiserum. The results are rocket-like precipitation lines, whose heights are proportional to antigen concentration. A standard curve can therefore be made, similar to that in Figure 5.15c.

Preparative Immunoprecipitation Immunoprecipitation can also be employed to isolate a protein from a complex protein mixture (serum, cell-extract, culture supernatant) if a specific antibody is available. The protein mixture should not be too concentrated, in order not to sediment too many irrelevant proteins during precipitation. A concentration of 1–5 mg ml^{-1} protein can be given as a guide. The protein is diluted with a physiological buffer, pH 7.2 (PBS, phosphate-buffered saline solution), and the antibody is added in small portions, mixed, and allowed to saturate the antigen slowly in order to prevent as much as possible the formation of soluble immune complexes through antibody excess (Figure 5.6). After incubation of the mixture at 4 °C for some hours or overnight, the precipitate is centrifuged down at 13 000 rpm for 30 min and after polyacrylamide gel electrophoresis the gel-selected protein can be immunochemically or chemically (amino acid sequence analysis) characterized. If a polyclonal

antibody is involved that is induced with the help of complete proteins, one can expect a precipitate with multivalent binding. In contrast, peptide antibodies or monoclonal antibodies generally lead only to soluble immune complexes if no repetitive epitopes are present (Figure 5.7), which is normally the exception.

These soluble immune complexes can then in cases of missing precipitation be precipitated at the equivalence point with a second antibody that is directed against the first in order to obtain a complete precipitation (Figure 5.6). The equivalence point can be determined in a simple way by mixing the primary antibody with different amounts of the secondary one in a series of tests. The heaviest precipitation indicates the point of equivalence. One places in each of six Eppendorf tubes 30 µl of the first antibody and then adds in increasing amounts 3, 6, 12 . . . up to 96 µl of the second antibody. These are mixed and incubated at room temperature for 30–60 min. Then the resulting precipitate is sedimented and from this precipitate the first impression of the equivalence point is gained. For more accurate determination one can investigate the clear supernatant by means of immunodiffusion (Figure 5.10) for the presence of first or second antibody. For this one applies, in agarose gel (Figure 5.10; Figure 5.11), each of the six supernatants against the first and then the second antibody and so finds the answer. If one finds antigen (first antibody) one is in the zone of antigen excess, and in the tubes in which the antibody (second antibody) is present one is in the zone of antibody excess. With the absence of antigen and antibody in supernatant and the finding of a maximal precipitate, one has identified the point of equivalence.

The behavior of the amounts of reagents can now be used for the optimal precipitation. In cases where the soluble immune complex is found in too low concentration (as with the use of radioactive reagents) one uses larger amounts of an additional "carrier" precipitating immune system without radioactivity and adds both partners (first and second antibody) one after the other in an amount at equivalence to the antibody with bound radioactive antigen. The carrier immune system leads, due to its protein amount, to a visible precipitate that can be handled. The sought after radioactive antigen with the bound antibody is then also found in this carrier precipitate.

Membrane proteins or hydrophobic proteins (apolipoproteins) remain lipid-bound in PBS and must therefore be separated and rendered monomeric before precipitation with the help of non-ionic (rarely ionic) detergents before they can be precipitated and isolated in the above-mentioned ways. Since with lipophilic proteins rather non-specific interactions occur than with hydrophilic proteins, a similar protein mix without antigen should be precipitated and analyzed in the same way, to identify the respective false positive interactions so that they can be subtracted. To avoid the presence of antibody in the precipitate, the first antibody can be bound covalently to a small corpuscular carrier or bead. After binding the immune complex (Box 5.1) the impurities can simply be washed away by centrifugation. If the carrier contains iron, the unspecific antibody can be separated from the remaining solution with a magnet. If the carrier is larger and if a column is used for separation, one speaks of affinity chromatography (Section 5.3.3).

Radioimmunoassays (RIA) The most sensitive method by which to measure antigen concentrations also in complex protein mixtures is represented by radioimmunoassay (RIA), published first in 1960 by Rosalyn Yalow and Solomon A. Berson (Nobel Prize in Physiology or Medicine for R. Yalow, 1977). Using this method, antigen concentrations of 0.5 pg ml^{-1} can be determined. The competitive radioimmunoassay uses the antigen–antibody reaction in solution after mixing radioactively labeled (mostly ^{125}I, ^{3}H) soluble ("hot") antigens with an increasing amount of unlabeled ("cold") antigen. By this method the radioactively measured "hot" antigen is replaced with increasing concentration of "cold" antigen. The displacement can be calibrated via known amounts of antigen. From the resulting standard curve, an unknown amount of antigen can thus be measured (Figure 5.17).

The exact measurement of bound amounts of antigen requires the separation from unbound antigen. This can be done in various ways, for example, by precipitation of the antibody. With this, the unbound antigen must nevertheless remain in solution. This simple precipitation of antibodies occurs, for example, in 35–50% saturated ammonium sulfate solution or by adsorption of the immune complex to bentonite or activated charcoal. If, under these conditions, the antigen also completely or partly precipitates, one can precipitate the total amount of antibody by means of a specific immunoprecipitation. Since the equivalence point for maximal precipitation is difficult to achieve owing to the low antibody concentrations, one often submits an optimal precipitating carrier antigen–antibody system in excess, which includes the minor

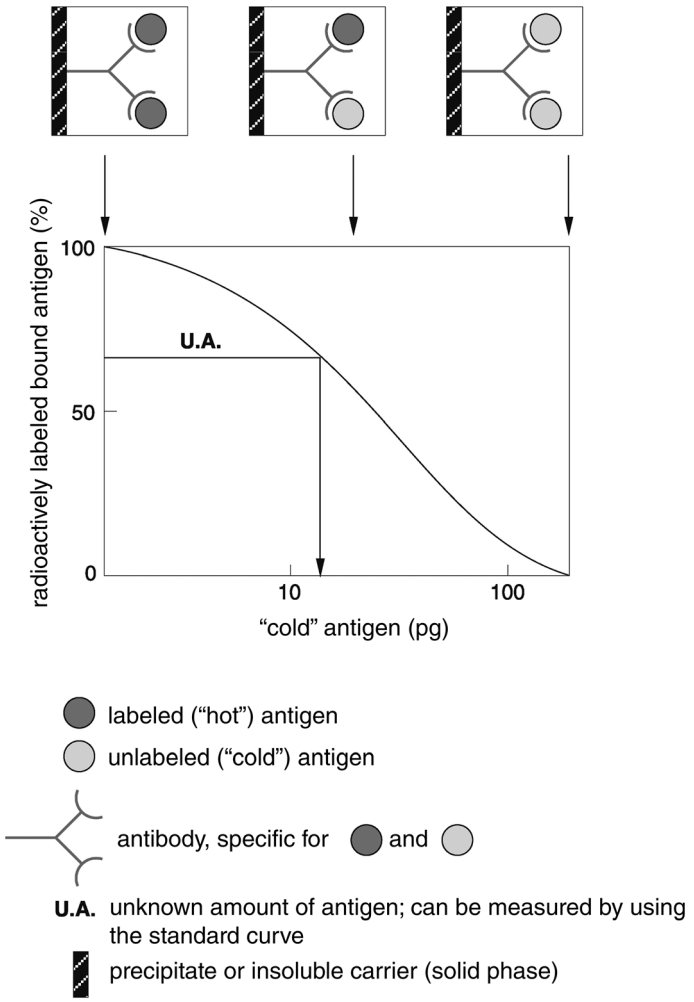


Figure 5.17 Competitive radioimmunoassay (RIA) in solution or on the solid phase for quantification of soluble antigens.

antigen–antibody complex that needs to be precipitated. In this way the small amount of the analytic antibody is easily co-precipitated (see above).

One can also couple the analytic antibody to an insoluble carrier and so remove the bound antigen from the unbound by washing. Additionally, the total antibody including the antigen-binding antibody can be separated from the unbound antigen by means of an immobilized secondary antibody. Variants of RIA in which one partner is immobilized are called solid-phase RIA and will be discussed in more detail in Section 5.3.3.

Numerous different approaches utilize the extremely high sensitivity and specificity of RIA for the most varied applications in all areas of biosciences, especially in medicine. Routine tests for the most varied plasma components with low concentrations such as proteohormones (ACTH 2 pg.ml^{-1}; insulin 5 pg ml^{-1}; calcitonin 10 pg ml^{-1}), steroid hormones (testosterone 50 pg ml^{-1}; progesterone 20 pg ml^{-1}), interleukins, enzymes, antibodies, pharmaceuticals (digoxin 100 ng ml^{-1}), drugs (morphine 100 pg ml^{-1}), or viral and respectively bacterial products are mentioned here, simply to underline the importance of RIA.

Despite its great versatility and enormously precise measurement accuracy, RIA has severe disadvantages, which are associated with the radioactivity essential for its sensitivity. The lack of storability owing to the half-life of radionuclides, the transport difficulties, and all the associated regulations for the use of radioactivity, with regard to protection of personnel, high costs for the building and maintenance of controlled areas where radioactivity is permitted to be used, the expensive counting equipment, and the financial problems concerning disposal of radioactive refuse are mentioned here. Alternative methods were therefore developed that avoid these disadvantages (see below, Enzyme immunoassay).

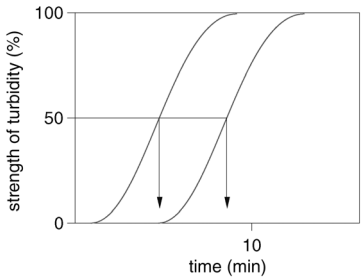


Figure 5.18 Nephelometry: The turbidity measurement indicates the antigen concentration from the speed and strength of turbidity. The increase is measured here at the inflection point of turbidity.

Nephelometry In recent decades nephelometry, which is based on immunoprecipitation, has come to the fore. This method quantifies the antigen concentration via measurement of the light scattering induced by an immune complex formation in solution. Since the kinetics of immune complex-induced clouding follow the rules of the Heidelberger curve (Section 5.3.2; Figure 5.6), the degree of cloudiness only correlates with the ascending leg of the curve, which means in antibody excess, with the amount of the antigen. To be certain that the measurement is performed in antibody excess, several dilution steps of the antigen should be applied for each given antibody concentration. Since the cloudiness of protein solutions can have manifold causes, it is important to note and take account of initial turbidities and others that non-specifically arise during the mixing of complex protein solutions with the corresponding buffers or with a pre-immune serum (an antiserum before immunization without specific antibody) and make allowances for them in the evaluation. According to the type of antigen, the speed of clouding or the maximal turbidity reached is measured and later converted by means of the standard curves. One can also determine the time to reach the inflexion point or the half-maximum turbidity value (Figure 5.18).

In turbidity measurements the total strength of the interaction, the avidity, and the affinity (Section 5.3) are included. Since the antigen–antibody reaction follows the law of mass action, the speed of reaction at constant antibody concentration depends solely on the concentration of the antigen.

The measurement time of about 5–12 min and the automation possibilities have led to nephelometry becoming the most important method in routine medical practice for the immunological quantification of proteins in serum, urine, and other liquids with complex protein mixtures. There are commercial automats that can determine the concentration of different serum proteins simultaneously in short times. The sensitivity of nephelometry is about 20 µg ml^{-1} and is thus considerably less than that of RIA and also somewhat less than that of enzyme immunoassays (Section 5.3.3). The increase of sensitivity by about tenfold is achieved by turbidity measurement of immune reactions on small corpuscles.

5.3.3 Immune Binding

Owing to the high sensitivity and the extraordinary versatility, binding tests today play a pre-eminent role in biosciences and medicine.

If one of the partners of an immune reaction is immobilized, that is, either the antigen or the antibody, then assays involving the corresponding antigen–antibody reaction are known as binding tests. Immobilization is mostly achieved by adsorption to a protein-binding plastic surface (e.g., polystyrene) or on an insoluble carrier through covalent coupling.

Clearly, an antigen–antibody reaction at an insoluble carrier is not immediately visible. For some years it has been possible to measure an antigen–antibody binding in real time with high sensitivity equipment under certain requirements on a standardized carrier, for example, through plasmon resonance (Biacore, Figure 5.25).

To detect the antigen–antibody reaction on an insoluble carrier, visible amplification systems are employed, such that the amplification can be done through fluorescence, luminescence, radioactivity, reduction of silver salts, enzymatic color reactions, or electron-dense corpuscles. Another important advantage is that the immune detection aided by binding tests is already possible with a single epitope–paratope binding, without the necessity of a bi- (or multi-)valent crosslinking, which is essential for visualization of a successful immune reaction involving agglutination or precipitation. Since monoclonal antibodies (Section 5.7.1, Monoclonal Antibodies) react solely with a single epitope, binding tests are ideal for their use. Nevertheless, the difference between specifically bound and non-specifically adsorbed antibodies must be regarded, which poses the central problem in all binding tests.

Because proteins generally adhere to almost every possible material, non-specific binding of the reaction partner (antigen or antibody) to the immobilized carrier can occur. It is necessary to minimize or prevent this *non-specific* adsorption and to strengthen the *specific* adsorption so that the reaction distance between the two bindings is maximized. If, in the case of maximal specific reaction, non-specificity approaches zero, then the system is optimized and the specific (or diagnostic) window has been found. In this case the system is ready for use, as illustrated in Figure 5.19.

While, in the case of immunoprecipitation methods, the complex process of formation of the antigen–antibody complex as an insoluble spatial network (Figure 5.6b(ii)) guarantees the high specificity of this method, with all immune binding tests, because of the always-present, competing, and multiple non-specific adsorptions, *searching for the specificity window,* that is, the condition of the epitope–paratope equivalent for each newly introduced test system, is obligatory because, for the stated reasons, immune binding tests rarely measure specifically at the first attempt since the concentration of the insoluble partner is already fixed.

With the setting up of a binding test one should widen the distance between specificity (signal) and non-specificity (background). With the application of an antiserum with high titer and higher affinity the antiserum can be further diluted while conserving the reaction strength.

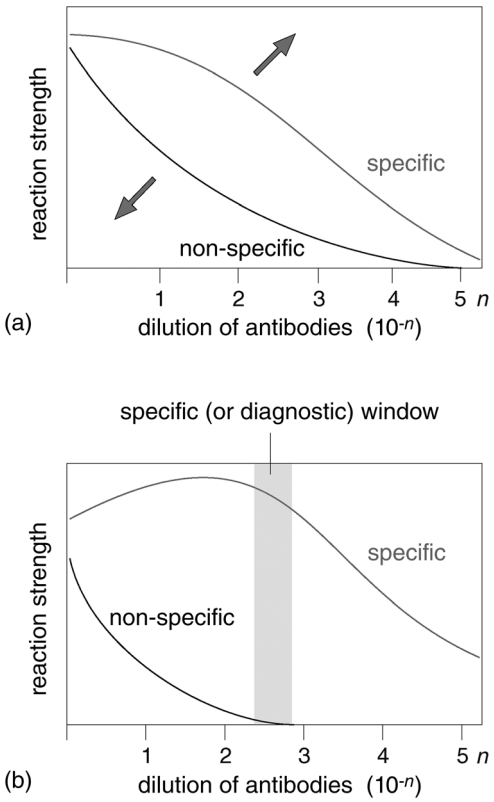


Figure 5.19 Binding tests: Optimizing and discovery of the specific window. (a) Not optimized: The difference between specific and non-specific reaction is too small for an exact measurement of a specific reaction. The arrows give the goal of optimizing, which consists of enlarging the distance between unspecific and specific reaction. (b) Optimized. The distance between specific and unspecific binding is so enlarged that an optimal measurement range allows exact measurements within the specific (or diagnostic) window.

With this dilution, the unspecifically associated proteins and the controls are diluted as well, which leads to a greater separation between signal and background and increased sensitivity. Additionally, the choice of a more sensitive amplification system (see Figure 5.20 below) can improve the signal to noise ratio. After optimal fixing of one of the reaction partners (antigen or antibody) by adsorption or covalent coupling to an insoluble substrate (nitrocellulose, glass-fiber, nylon, or other support), the non-specific adsorption can be prevented by blocking the remaining binding positions with the help of gelatin, bovine serum albumin, casein, skim milk powder, egg-white, or other proteins, also protein mixtures.

For the inhibition of non-specific interactions, the use of certain non-ionic detergents in the wash solution is time-proven. Other non-specific results owing to interaction between blocking agents and reagents or between reagents themselves must first be absorbed in in solution or blocked before their use.

All binding tests require, for rendering the expired immune reaction visible, an amplification system that allows qualitative detection and quantitative measurement. Numerous different systems are known. Some examples of important arrangements for visualization are discussed here and shown in Figure 5.20.

Visualization of an immune reaction relies on a stable bridge between the sought reaction partner (antigen or antibody) and the amplification system (A in Figure 5.20). Amplifiers (or indicators) that have been used include fluorescent dyes (Albert H. Coons introduced in 1941 fluorescein isothiocyanate and with it founded immunofluorescence microscopy), luminescent dyes, radioactivity, enzymes (Figure 5.20a–f and i–l, Figure 5.23a below), reduction of noble metal salts and electron-dense corpuscles like colloidal gold (Figure 5.20g, h, Figure 5.23b below). The amplifiers are covalently or non-covalently (adsorptively) coupled to the reactants.

Serving as stable bridges between the indicator and the sought reaction partner are the bindings: antigen–antibody (Figure 5.20b–d), biotin–avidin (Figure 5.20e), protein A (Figure 5.20f–h), lectin, (Figure 5.20i), and antigen (Figure 5.20j–l). A bifunctional antibody (directed against two different antigens) is shown in Figure 5.20j. The techniques shown here work also with an antibody with two identical paratopes, so that the antigen bound to the solid phase is the same as the labeled one. Antigens can be identified with systems (a) to (j) and antibodies with the systems shown in (k) and (l).

Fluorescent Dyes, Table 6.1

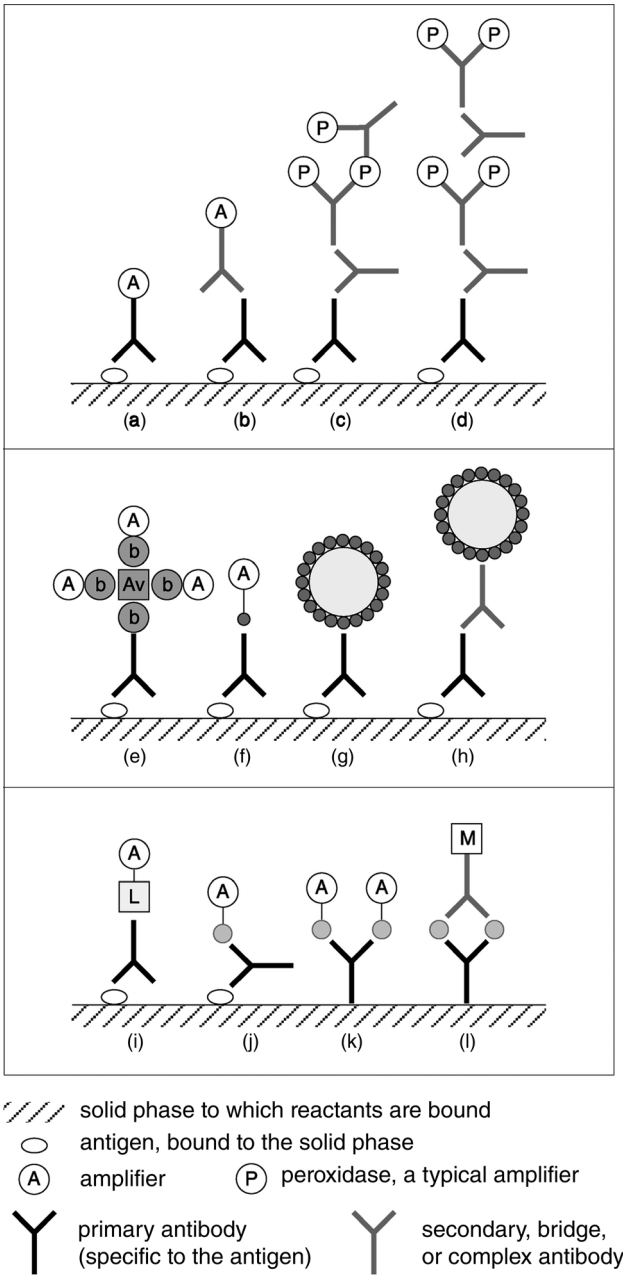


Figure 5.20 Binding tests: Making visible the antibody binding. (a) Direct method; (b) indirect method; (c) unlabeled or PAP (peroxidase–anti-peroxidase) method of Sternberger 1970 (compare Figure 5.23a), peroxidase (P) serves as amplifier (A) here. Shown is a polyclonal PAP-complex. (d) Double unlabeled method. Double-PAP-method, shown here with a monoclonal PAP-complex (compare Figure 5.23a). (e) Avidin–biotin complex (ABC-technique), direct. More usual is the indirect ABC technique, in which the secondary antibody is biotinylated. This system can be further amplified. (f) Identification using amplifier-labeled protein A. (g) Direct protein-A–gold method. (h) Indirect protein-A–gold method (compare Figure 5.23b). (i) Identification using amplifier-labeled lectin. (j) Identification using labeled antigen. (k) Identification of the specificity of antibodies in cells and tissues by means of labeled antigen. (l) Identification of the specificity of antibodies in cells and tissues by means of the Sandwich method. The antibody binds the homologous antigen, which can be identified subsequently using one of the amplifying systems (M = bridge and amplifier (A), (a–i)). For use in solid phase radioimmunoassay or in micro-ELISA the antibody serves as antigen concentrator (catching antibody).

In the direct arrangement (Figure 5.20a), the primary antibody is always covalently coupled with the amplifier. With this arrangement the testing run is short. If the already coupled (i.e., labeled or tagged) antibody is not commercially available, it can first be isolated with affinity chromatography and subsequently coupled with the amplifier. The affinity chromatographic isolation of the pure antibody assumes, however, the availability of the isolated antigen and a large amount of specific antiserum for each antigen to be analyzed.

The indirect arrangement (Figure 5.20b) circumvents this time-consuming and financially expensive method of labeling the respective primary antibody by the use of an anti-immunoglobulin against the primary antibody, which is tagged with the amplifier. This secondary antibody, which is won from another animal species, binds to all primary antibodies of any given specificity from the first species. A further advantage of the indirect method is higher sensitivity, since the second antibody can mean an amplification of about twentyfold to hundreds-fold.

Affinity Chromatography, Section 10.4.8

An even more sensitive method is the unlabeled method of Sternberger in which three antibodies are applied, one after the other. As shown in Figure 5.20c, the primary antibody is bound with a corresponding secondary antibody in excess, so that, statistically, one monovalent binding remains free, which can then bind a third, against an amplifier directed antibody, which must be won from the same animal species as the primary antibody. This third antibody is labeled with an amplifier or indicator.

When horse-radish peroxidase (P) is used as indicator enzyme, the amplifier is known as a peroxidase–anti-peroxidase or PAP complex. Peroxidase splits peroxide, liberating highly-reactive oxygen radicals that can themselves oxidize certain chromogens. The resulting color development is visible and quantitative. The PAP system is approximately over 1000 times more sensitive than the direct system.

To achieve a further increase in sensitivity, the secondary antibody (also known as the connecting antibody) and the PAP complex can be applied a second time (Figure 5.20d). Admittedly, with this increase in sensitivity the amount of non-specificity can also increase.

The polyclonal PAP-complex consists of three peroxidase molecules and two antibodies (Figure 5.20c), while the monoclonal consists of one antibody and two peroxidase molecules (Figure 5.20d).

The bound primary antibody can be made visible not only via the antigen–antibody reaction but also with other binding systems. A highly specific system with extremely high sensitivity is the extraordinarily stable avidin–biotin complex formation (ABC system; Figure 5.20e) where the primary (shown here) or the secondary antibody can be biotinylated. The amplifiers (A), peroxidase or the alkaline phosphatase (AP) are biotinylated and become bound to the tetravalent avidin with saturation of three binding positions. The remaining binding positions can then bind to the biotinylated primary (shown here) or the secondary antibody (indirect ABC method). If the amplifiers are multiply biotinylated, very large avidin-–enzyme complexes are produced, which can lead to a strongly increased sensitivity of the ABC method.

Another type of binding can result from bacterial proteins, like the bacterial protein A (Figure 5.20f–h) or protein G, which bind selectively to the Fc part of certain IgG isotypes of humans and animals.

The antibody binding can be recognized either via labeling with corpuscles or at primary or secondary antibodies. In such cases one uses erythrocytes or synthetic microbeads as visible indicators for light microscopy; in addition, ferritin, virus particles, or colloidal gold particles can be used for electron microscopy. Colloidal gold particles can adsorb protein A (Figure 5.20g and h), antibodies, or other proteins and can thus provide the bridge to antigen or antibody. In addition, lectins and even antigens can be used as bridges between amplifier and antigen, as depicted in Figure 5.20i–l. With the help of labeled antigens one can identify even antigens (Figure 5.20j) as well as antibody specificities (Figure 5.20k and l). The last two arrangements can also be used to detect immunohistochemical antibodies of sought specificities in tissue. A particularly interesting arrangement is the sandwich arrangement (Figure 5.20l) as a (mechanically) fixed antibody can bind soluble antigen before this is identified using an antigen detection system (M). The antigen binding via a fixed antibody leads to an antigen concentration, which gives very strong amplification and thus allows the identification of very low concentrations of antigen (see below). Other amplifiers are radionuclides (see radioimmunoassay, above); these, however, because of environmental considerations have been superseded by other amplifiers like enzymes (Figure 5.20) whose mode of operation with catalytic properties is precisely measurable. Enzymes are storable, environmentally friendly, and cheap.

Highly sensitive identification of a protein is possible using immunodetection. If a protein is present in a concentration of about 0.1 mg ml^{-1} and from this solution 0.2 µl (total therefore of 20 ng protein) is dropped on a nitrocellulose membrane and the membrane first blocked and then investigated using one of the above-mentioned detection systems (e.g., the PAP method), a strong specific coloration of the protein spot occurs, with which the protein can be identified specifically. The sensitivity can be increased so far that protein amounts of <1 ng can still be recognized. If one applies a series of different protein concentrations, a calibration curve can be produced and the amount of unknown antigen can be determined (semi-)quantitatively.

The dot-immunoassay can be used for fast orientation. The test is very useful, especially for the specificity checking of antisera and the selection of monoclonal antibodies, since a large series of different antigens can be tested out simultaneously with the same antibody preparation.

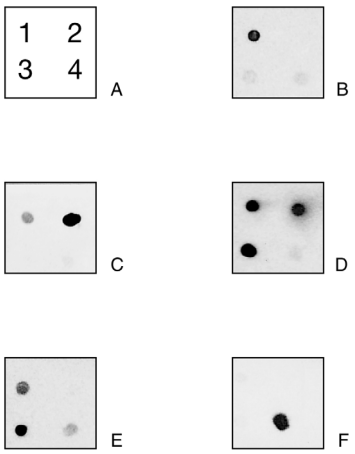


Figure 5.21 Dot immunoassay. The example shows an epitope mapping of monoclonal antibodies (A–F) that were produced against β_2-microglobulin (β_2m) and its fragments. Antigens numbered 1–4 were dropped at a concentration of 0.05 mg.ml^{-1} on to a nitrocellulose membrane and the membrane was then blocked with a gelatin solution. The binding of the primary antibody was made visible with the PAP method. Antigens 1–4 are β_2m and synthetic peptides (the numbers correspond to the position of amino acids of β_2m). 1 β_2m, intact, 1–99; 2 β_2m, 1–19; 3 β_2m, 9–24; 4 β_2m, 20–36. The monoclonal antibodies B–F react quite differently, as shown in this orientation test.

Figure 5.21 shows how a dot-immunoassay can be utilized for epitope mapping of monoclonal antibodies. Shown is the binding of monoclonal antibodies against different regions of the native serum protein β_2-microglobulin (β_2m) and with the use of corresponding synthetic peptides. One perceives that the monoclonal antibodies bind at completely different regions of the β_2m. The monoclonal antibody B binds with an epitope on the native protein, while F only reacts with peptide 4, and not, however, with the native protein. Perhaps F recognizes a hidden epitope of β2m (Section 5.2). On the other hand, the antibodies C–E bind to various epitopes, as determinable by the peptide sequence (Figure 5.21) that also lie exposed in the native protein and are thus accessible through the antibodies C–E (Figure 5.5).

Western Blotting, Protein Transfer, Immobilization, and Immunodetection Protein blotting with subsequent immunodetection was introduced by Harry Towbin in 1979.

The usual trivial name, "western blotting", goes back to the name of the discoverer of blotting techniques, Edwin Southern, who introduced in 1971 a method for the fractionation of DNA fragments and subsequent hybridization, so-called "Southern blotting". In reference to his name, the corresponding fractionation of RNA fragments was named "Northern blotting" and protein blotting became Western blotting.

Actually, the term western blotting describes only protein blotting with subsequent immunodetection, but often it is used to mean protein blotting in general including the modification electroblotting. The principles of dot immunoassay follow the technique of western blotting except in the latter case the soluble antigen is not immobilized by dropping from a pipette on to a carrier but instead the proteins are fractionated in a gel, and then transferred by capillary (Figure 5.22a) or electrotransfer (Figure 5.22b) onto a carrier (e.g., nitrocellulose) and immobilized for the subsequent immunodetection. It is important that during the transfer the proteins on the carrier appear in the same geometric order that they had after fractionation in the gel. As shown in Figure 5.22b, the sought protein is identified on the membrane, as in dot immunoassay, by means of the different or similar detection systems, shown in Figure 5.20.

Electroblotting, Section 11.7

In Figure 5.22b(i1) one discerns the fractionation of a protein mixture, stained with Coomassie blue. In Figure 5.22b(i2) an isolated protein from the protein mixture is fractionated in the same way and stained. One sees a uniform protein. After electrotransfer on to a nitrocellulose membrane and subsequent immunodetection, only one band is seen in the protein mixture.

This application method, used extensively in the biosciences, is extremely versatile since proteins can be fractionated and immunodetected that are insoluble under physiological conditions. For example; membrane proteins in urea and ionic detergents can, after successful fractionation, after electrotransfer and immobilization on a membrane, be transferred in a compatible buffer for immunoidentification and then further investigated.

In a series of biological systems the amount of different proteins can be so great that the individual constituents cannot be sufficiently separated in one dimension. In this case, one can increase the resolution by the use of two-dimensional fractionation, such that the first dimension fractionates by electric charge (e.g., by isoelectric focusing) and the second dimension incorporates separation by size.

2D-Electrophoresis, Section 11.6

The results are clouds of points with a resolution of up to 10 000 proteins per gel plate (Figure 5.22b(ii)). The high resolution is mainly used to show whether a protein is homogenous with respect to charge *and* size.

Figure 5.22b(ii) gives an example of an homogenous protein (1). In comparison, protein (2) is uniform in size but not in charge, because in the separation by charge it separates into three spots (charge variants). In addition, protein bands of homogenous charge (3) can produce spots of different size classes, especially if the following size separation step is done in reducing buffer, such that the covalent structure of a protein is visible. After transfer to a nitrocellulose membrane the immunodetection of individual proteins ensues and after the transfer to a glass fiber membrane the N-terminal amino acid sequence can be established and the size or charge variants further examined.

Accordingly, one can contemplate, starting from the corresponding oligonucleotides, searching DNA- or cDNA-sequences for identification of the complete nucleotide and protein sequence, as far as the derived amino acid sequence allows.

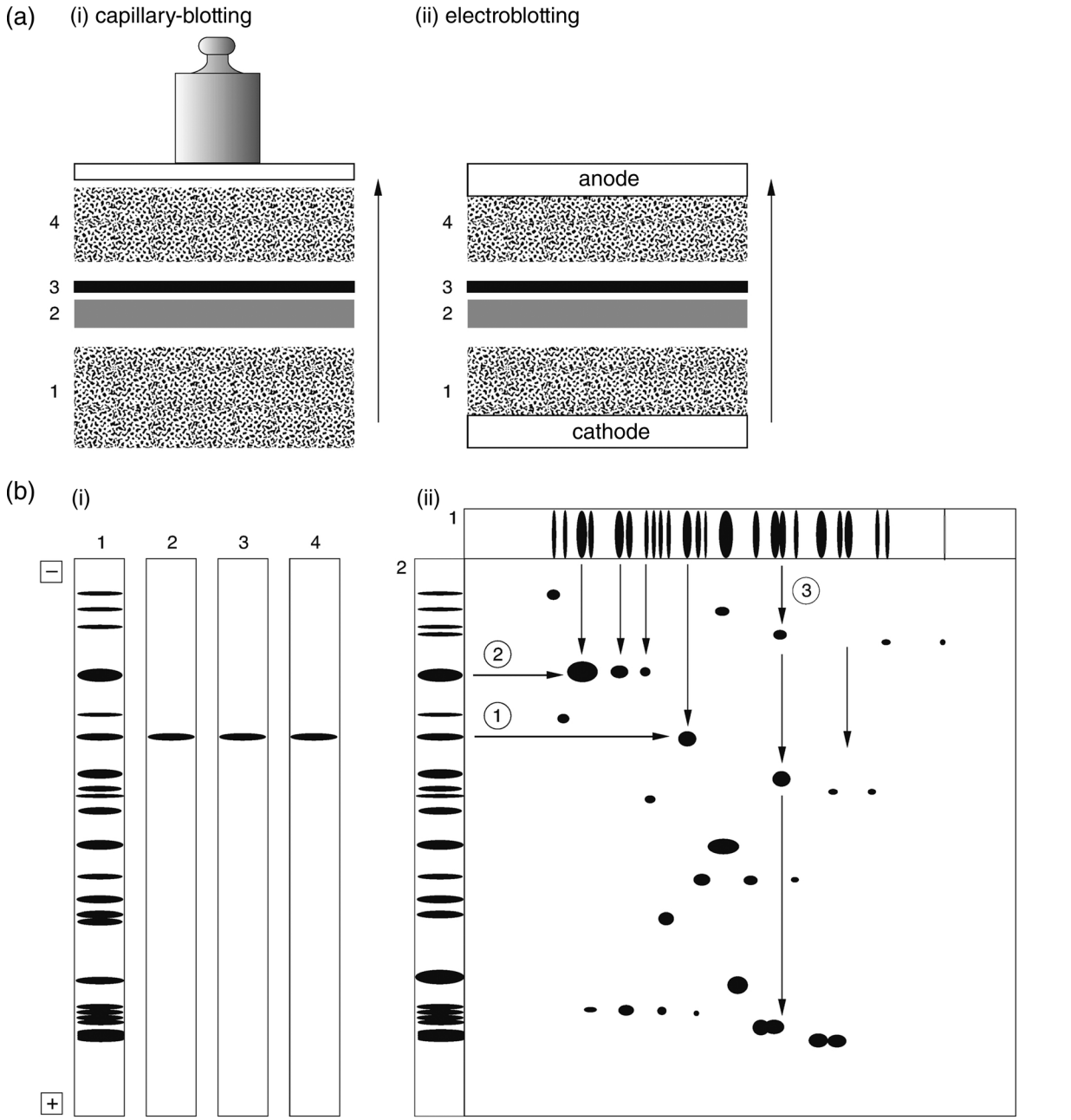


Figure 5.22 Western blotting: Protein transfer and immunodetection are shown schematically. (a) Protein transfer. (i) Capillary-blotting: (1) A filter-paper stack (that can also lie in a basin with buffer), saturated with transfer buffer; (2) gel with fractionated proteins; (3) carrier, usually nitrocellulose, on which the protein is transferred by capillary suction into the non-saturated filter paper stack in (4). The weight on the glass plate provides an even contact. The lower wet paper can be omitted if enough protein is available for immunodetection. (ii) Electroblotting: (1) Cathode with transfer buffer-saturated filter-papers; (2) and (3) as in capillary-blotting; (4) anode with filter paper, saturated as with (1). The electrophoretic transfer buffer normally contains 20% methanol. This protein transfer can take place submerged in buffer in a buffer vessel or semi-dry between filter papers with graphite or platinum-coated electrodes. The transferred and immobilized proteins can be investigated for their antigen content with the help of the previously-mentioned immunodetection methods (Figure 5.18).
(b) Results. (i) Western-blotting: (1) A protein mixture was fractionated by size in SDS in a polyacrylamide gel and the protein bands were stained with Coomassie blue (protein control); (2) an isolated protein was similarly separated and stained; proteins in (3) and (4), which correspond with those in (1) and (2), were stained by the indirect immunoperoxidase method. In the protein mixture only the sought protein gives an immunochemical single signal (3) or as a single protein (4), a demonstration of immunological specificity in both with 4 as the positive immunochemical control. (ii) Two-dimensional separation and Coomassie blue staining of a protein mixture of b(i1). (1) Separation according to charge by means of isoelectric focusing; (2) separation according to size. As reference a single-dimensional separation in b(i1) is provided. By means of the much higher resolution in the second dimension the inconsistency of single bands obtained in one-dimension is made visible.

Two-dimensional separation is equally important for the reverse process, which means that starting from DNA/RNA part-sequences the corresponding proteins and their variants can be identified using the immunodetection method (Figure 5.20) via the synthesis of immunogene peptides and the induction of peptide antibodies (Section 5.7) in two-dimensional blots.

Immunohistochemistry, Immunocytochemistry Whereas previous sections describe methods with whose help soluble or insoluble proteins can be investigated, this section is concerned with proteins *in situ,* at the locality of their production, their function, or their degradation. With the help of these methods, the question of microscopical or electron-microscopical localization of definite proteins and other substances within tissue complexes or within cells can be answered. Immunodetection in tissue (immunohistochemistry) and in single cells (immunocytochemistry) has become, since the introduction of immunofluorescence by Albert H. Coons 1941, an established standard method (with many variants) for the mostly qualitative recognition of certain antigens.

The preparation of tissues and cells is of critical importance for the success of immuno-detection, for the antigenic determinants of the sought antigen must be accessible in the use of these methods. Today there are light and electron optical immunochemical methods for the investigation of both native tissue sections and sections from fixed tissue that is embedded in paraffin or plastic.

Immune reactions can occur in the native tissue, that is, before cutting and embedding pre-processing. These methods were preferred in previous years because of the availability of antigenic determinants.

Basic problems are, however, the poor penetration of relatively large antibodies into the tissue, and the difficult washing out of the non-bound reagents. Today, immunodetection is mainly carried out after processing, on the section itself, which can be native or fixed. Many of the necessary antigen determinants can be kept unchanged with certain fixatives, for example, with 2% paraformaldehyde solution at 0 °C for several hours. On the other hand, one can partly reverse the fixation-related loss of antigen determinants by enzymatic pretreatment of fixed tissue or by etching the plastic sections with H_2O_2 or sodium ethoxide and many other methods known also as antigenic retrieval.

Today, mostly fixed tissue sections are used for immunohistochemistry because of the easier handling and better morphological preservation. Above all, it appears that discovery of the specific window with the use of a different concentration series of various reagents is much easier when the optimization is carried out directly on the tissue section after the processing. Since we are concerned here with classical binding tests, optimization of the detection system and the discovery of the specific window follow the rules given previously (Section 5.3.3).

In light microscopy one uses mostly the indirect method, PAP, or the ABC method to discover the binding of the primary antibody to the sought after antigen in tissue sections.

In contrast to chromogens, which are used in micro-ELISA (see below) and which remain soluble after oxidation of the chromogen, the soluble colorless chromogens used here become insoluble so that they stay attached as colored markers in the immediate vicinity of the antigen in the tissue.

At the electron-microscopical level electron-dense corpuscles have proved successful. Because of its great versatility, colloidal gold is often used, because it is electron-dense and produced in various sizes for the labeling different antigens on the same ultra-thin section.

An example of the detection of pathological protein deposits in paraffin sections of a human kidney using the PAP method (Figure 5.20c) is shown in Figure 5.23a and that of the same protein deposit in the electron microscope using the protein-A–gold method (Figure 5.20g) is shown in Figure 5.23b. Methods are known with which various antigens can be identified on the same section. In this case a differently colored chromogen is available for each antigen, giving improved differentiation.

Cell biology has had a great upsurge with the possibility of following the fate of myriads of proteins and protein complexes in both living cells and living tissues. Interest in cell and tissue development is concerned with ontogeny, function alterations during inflammation, apoptosis, and neoplasy. The problem of cell membrane penetration of antibodies into the living cell was more or less solved with the use of antibody-fragments (Fv-fragment, Figure 5.1), mini-antibodies. The latter can be selected from single-chain synthetic antibody-mimetic polypeptide libraries as fixed epitopes or won via the generation of monoclonal antibodies via antibody

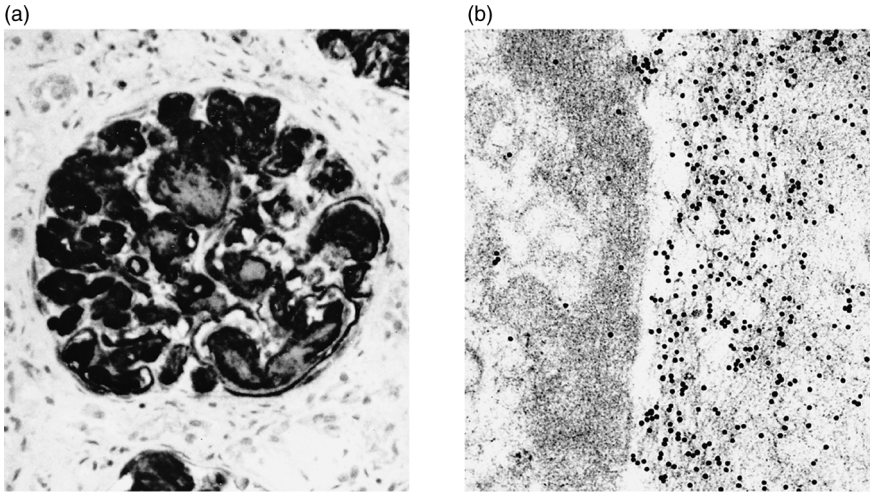


Figure 5.23 Example of immunohistochemical detection at light- and electron-optical levels. (a) Immunohistochemical detection of pathologically accumulated proteins (amyloid deposits) in a kidney glomerulus in a patient with nephrosis and rheumatoid arthritis. The protein deposits are diagnosed as AA-amyloidosis by means of the PAP-method (Figure 5.20c) and amino-ethylcarbazole as chromogen (after oxidation brown–red and insoluble) with the use of a monoclonal antibody against amyloid-A protein (380×). (b) The same protein deposits were visible in electron microscopy as fibrils and decorated here using also AA-monoclonal antibodies with the protein-A–gold method (Figure 5.20g). One can discern electron dense black gold particles (whose binding is specific) lying along the fibril. The diameter of the gold particles is 17 nm.

fragments from immunized camel-related animals (whose antibodies only possess the heavy chain). These small-molecular antibodies become fluorescently visible in the cell after labeling with fluorochromes. The corresponding fluorescing proteins can also be attached to these mini-antibodies or analogues by gene technology. The fluorescence color for use of simultaneous multiple investigations can be changed by the incorporation of unorthodox amino acids. The most recent advance uses transfection of one-chain (camelid) mini-antibodies, which are produced intracellularly by the cell itself and by this circumvents the need for cell membrane penetration of the antibody.

Affinity Chromatography After immobilization of one reaction partner, the other soluble partner can be bound by an immune reaction, washed and selectively removed by special solvents (Box 5.1). Thus, in this way, an antigen can be isolated by a covalently attached specific antibody and, the reverse, a specific antibody can be isolated by a covalent attached antigen.

A large number of commercially available carriers can be coupled that, according to the method of affinity chromatography, have different properties (dextran, agarose, silicate, and other carriers). The carriers differ also in regard to the chemical group with which the protein can be coupled (e.g., —NH_2, —COOH, —SH).

The isolation is normally realized by binding and elution from granular carriers in a chromatography column. The bound and washed-free partners are then mostly eluted in acid (Box 5.1) and, finally, neutralized.

Affinity chromatography can, under appropriate binding and elution conditions, also be carried out with substrate–enzyme-binding, nucleic acid and complementary-strand binding, and every other type of biospecific binding.

The versatility of this method and the normally very high cleaning efficiency attainable with biospecific bonding (100–1000× concentrations are possible in one step, depending on the starting situation) have made affinity chromatography one of the most important preparative methods in biochemistry.

Affinity Chromatography, Section 10.4.8

Above all, the biospecific concentration of ligands can, with an extremely low concentration of proteins in highly complex protein solutions (and the hopelessness of isolation with the use of classical methods), be the decisive advantage.

Enzyme Immunoassays (EIA, ELISA) The development of the principle of enzyme-immune-assay (EIA) was first demonstrated by Eva Engvall and Peter Perlman in 1971 and named ELISA (enzyme-linked immuno-sorbent assay). It has, because of the necessary avoidance of the radioactivity used in RIA, become a leading technique.

EIA/ELISA follows the principles that have already been discussed concerning radio-immunoassays, except that the amplifier is not radioactive decay but instead an enzyme-catalyzed substrate with subsequent chromogen conversion.

The soluble and colorless chromogen is changed to a soluble, colored, and quantifiable dye, in contrast to immunohistochemistry, where the resulting insoluble dye remains held in the tissue. All EIA/ELISA reactions occur with an immobilized partner, which considerably simplifies the separation of bound and unbound reagents.

ELISA, which was originally performed in plastic tubes, is now known retrospectively as macro-ELISA, in contrast to micro-ELISA, which is performed in purpose-made 96-hole microtiter plates. Plastic surfaces actually do not bind antigens or antibodies covalently, but nevertheless bind with astounding tenacity. The greatest advantages are the relative simplicity of performing tests, the possibility of automation, and the measurement of color intensity of the chromogens in the shortest time by means of a special photometer, the micro-ELISA reader. Following this, the reading can be made by a computer program after calculation of a standardized curve.

One can distinguish basically between three ELISA systems, which exist in many variants, developed for all possible uses (Figure 5.24). Firstly, there is an EIA-system that (by using an enzyme-labeled antigen) can measure competing non-labeled antigens in complex protein mixtures (Figure 5.24a). This principle corresponds to competitive solid-phase RIA. Like RIA, the competition of the labeled antigen can be calibrated using standard solutions of unlabeled antigen. The amount of competition from an unknown antigen can be exactly quantified by means of a standard curve (Figure 5.24a(i)–(iii)).

A second variant of EIA, non-competitive ELISA, works with successively applied reagents (Figure 5.24b(i)). This system corresponds to the procedure of a typical binding test, in that antigen adsorbed to a microtiter-plate binds an antibody, which can be detected by means of typical amplification (A). The recognition and amplification system for binding detection is mostly indirect, but can also be another system, as depicted in Figure 5.20.

This ELISA variant is mainly used for studying antibodies, determining their titer or the disclosure of antigen or antibody specificities, as well as epitope-mapping of monoclonal antibodies. This system is, however, also suitable in general for investigating every other binding function not based on an antigen–antibody binding.

A further system, the "sandwich ELISA" (Figure 5.24b(ii)), serves mainly for quantification of antigens present at low concentration in complex mixtures. This ELISA begins with the immobilization of a specific antibody in a microtiter plate, known as the catching antibody, which binds the relevant soluble antigen, so that all other non-bound mixture constituents can be washed away.

This capture by antigen-binding corresponds to the binding step in affinity-chromatography. Especially important is the concentration increase of the antigen through the catching antibody. Through this concentration from greatly diluted solutions, sandwich ELISA is capable of attaining a sensitivity approaching that of RIA. The antigen bound to the catching antibody can now be determined in the same way as with simple non-competitive ELISA (Figure 5.24b(i)).

Since the epitopes of the catching antibody are already occupied, it is necessary to use a second labeled antibody with another epitope specificity for the detection of the bound antigen. This applies especially with use of monoclonal or other mono-epitope-specific antibodies.

In contrast to the discussed variants of EIA or ELISA, which are denoted as *heterogeneous* because of the multiple overlayering with diverse reagents, there also exists *homogenous* EIA. This variant suffices with a single tube and can be carried out in minutes, in contrast to heterogeneous systems, whose operation takes several hours.

Because the homogenous system includes no antibody-based amplification steps it is naturally less sensitive than the heterogeneous EIAs. Homogenous immune assays work according to the scheme shown in Figure 5.24c. The principle of homogenous EIAs is based on a change in an enzymatic activity, which is triggered by the antigen–antibody reaction. Homogenous immune assays are so constructed that either a catalytic center is covered and loses its activity through the binding of a nearby hapten (Figure 5.24c(i)) or an antigen can experience a steric conformation change through immune binding which leads to activation of a catalytic center (Figure 5.24c(ii)). Both resulting changes of enzyme activity can be quantified.

Through these changes in enzyme activity, therefore, an immune reaction can be recognized in the same tube immediately after addition of the antibody and be quantified. This *one-pot test,* whose preparation is not trivial, can easily be automated and enables extremely high sample throughput within a short time.

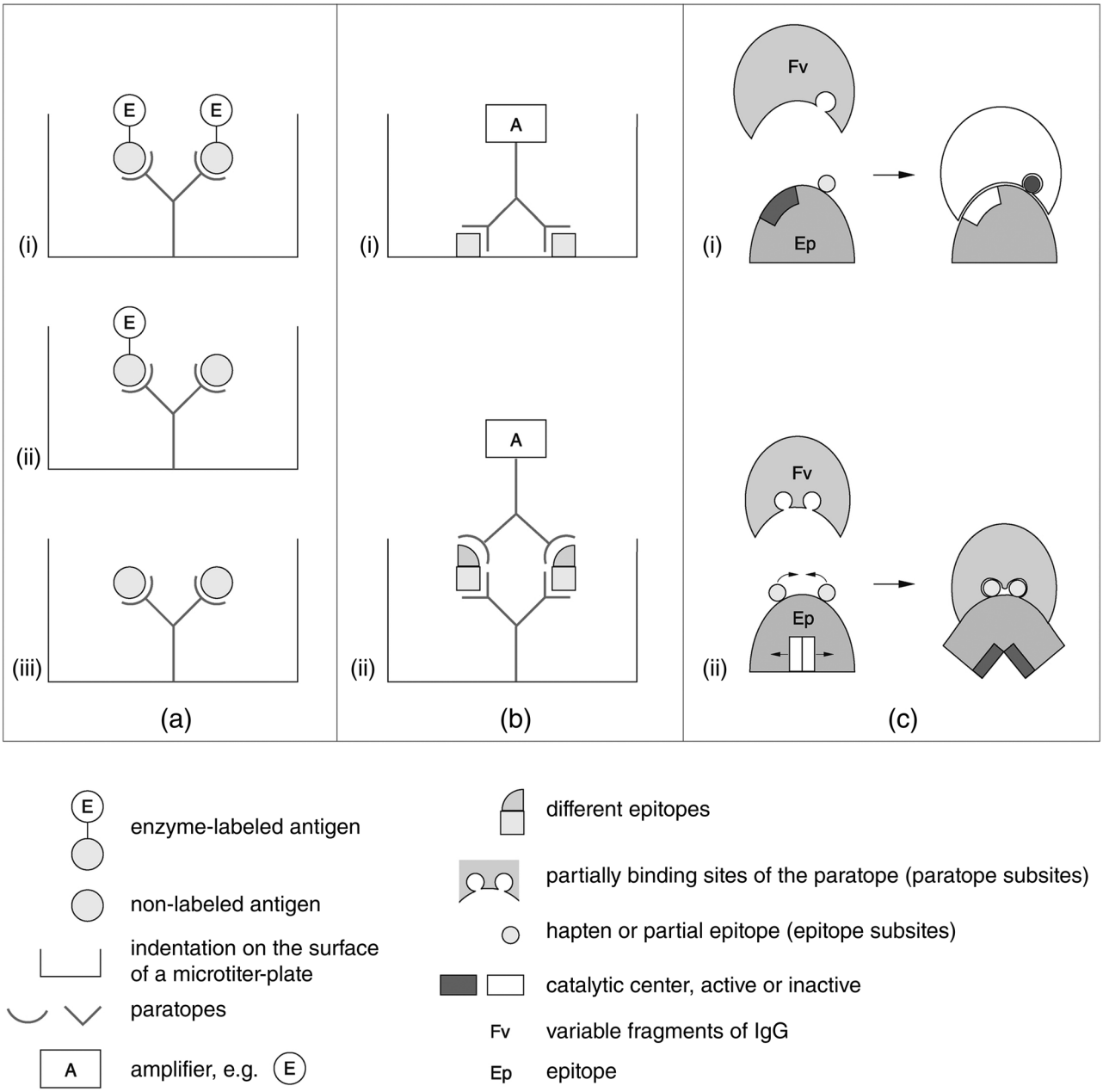


Figure 5.24 Enzyme immune assay EIA, heterogeneous and homogenous. (a) Competitive (like RIA, Figure 5.17), (i) with 100% labeled antigen, (ii) 50% displacement of E-labeled by unlabeled antigen, (iii) complete displacement of the labeled antigen by unlabeled antigen. E: enzymatic activity as amplifier. (b) ELISA: (i) simple, (ii) sandwich-ELISA; (A) amplifier as shown in Figure 5.20; the antigen is bimodally-bound by two antibodies of different specificities via two correspondingly different epitopes. (c) Homogenous EIA, schematic only. (i) Antigen–antibody reaction conditioned partial or complete inactivation of the catalytic center. (ii) Antigen–antibody-reaction conditional through a conformational change activating the catalytic center.

***Biacore* Technique** Biacore is a trade name of the first firm to offer this system, which can precisely identify, follow exactly over time and simultaneously quantify the process of specific interaction of two binding partners, especially antigen–antibody reactions. Since one of the two biomolecular binding partners is always immobilized, the *Biacore technique* belongs to the binding tests and is subject to the laws according to which the specific window for measurement of specific interaction of two partners can be found (Figure 5.19). This system can document the association, or alternatively the dissociation, of both partners in real time. One also gains important information concerning the reaction kinetics (speed of association and characteristics

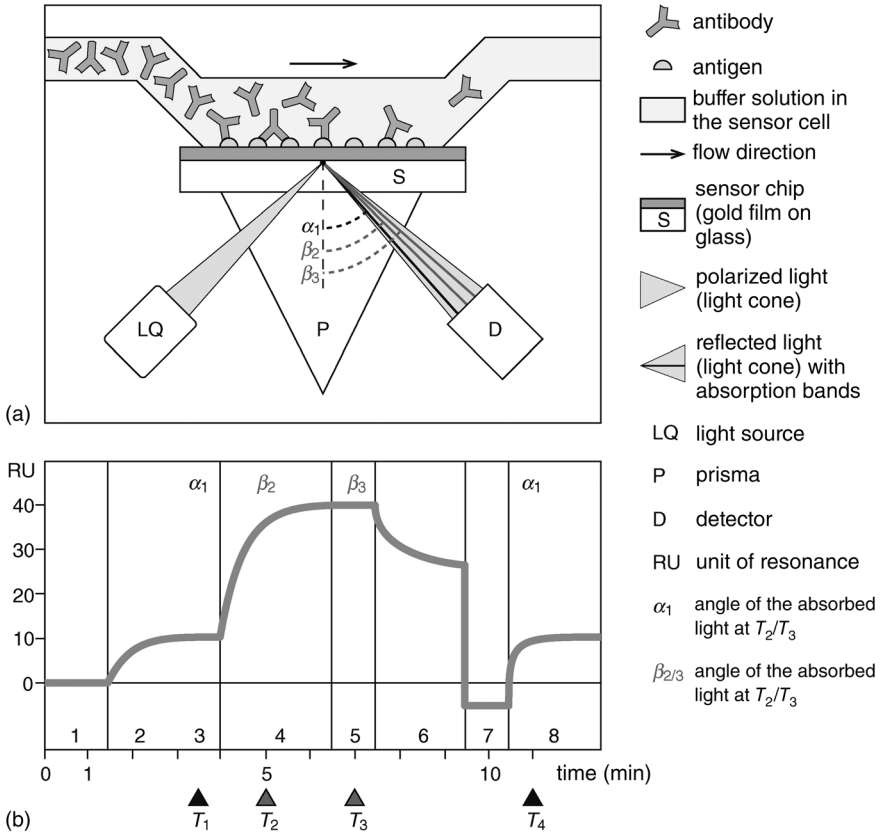


Figure 5.25 Measurement of antibody binding to an antigen using the SPR-Biacore technique in real-time. (a) Schematic description of the arrangement with light source (LQ), sensor chip (S), detector (D), and continuous flushable mini-chamber with insoluble antigen on the buffer-facing side of the gold film. (b) Measurement profile of the antigen–antibody reaction, which is depicted in (a). The profile shows, in phases 1–8, the amount of bound protein on the sensor chip. 1: Baseline (sensor chip without protein); 2: antigen adsorption on the gold film until saturation; 3: washing away of excess antigen; 4: increasing antibody binding; 5: antibody binding saturated; 6: buffer without antibody, dissociation, firstly of the low-affinity antibodies; 7: dissociation of the higher affinity antibodies with denaturing buffer (Box 5.1) and regeneration for a further test. α_1 Angle at time T_1, starting conditions (phase 1 and 8); β_2 angle at time T_2 (phase 4, of the stage depicted in (a)); β_3 angle at time T_3 in phase 5; RU, resonance units, one unit corresponds to 1 pg mm^{-2}. (The light cone passage through the prism is greatly simplified.)

of dissociation). The method is protective of the proteins – it works without labeling and requires miniscule amounts of protein. As well as this, the corresponding binding partner can be isolated in minutes, chromatographically, from complex mixtures by using correspondingly pure antigens or specific antibodies. The technique can also concentrate binding partners from highly diluted solutions. The method produces functional characteristics of proteins at the micro-level, a level that the miniaturization of protein analysis reached long ago (proteomics).

Biacore (and now several other firms with in part other principles, which also work free from labeling) meets the trend towards miniaturization by furnishing functional analyses at this level. The sensitivity – one can still measure a specific binding of 1 pg mm^{-2} – the speed and simplicity of operation following appropriate training (after finding the specific window) converge with the possibility of automation and *high-throughput* techniques.

The Biacore technique is based on a sensor chip, which can generate surface plasmon resonance (SPR). The SPR functions by using the change in reflection angle of polarized light on the glass inner side resulting from a plasmon of a gold film as a function of the amount of proteins located on the back side of the gold film (Figure 5.25a).

Surface Plasmon Resonance Spectroscopy, Section 16.6

This sensor chip, which is located in a continuously-flushable micro cell, can be exposed to various ligands and buffers intermittently and register the kinetics and binding profiles and separation profiles for each pair of binding macromolecules. Thereby, the change of SPR is converted into quantitative measurement values, as in the example depicted in Figure 5.25b.

This system can be used to discover reactants among a large series of possible substances, and also to the optimize the binding properties of isolated ligands, which, owing to the high throughput, is mainly used in the search for new medications from immunopharmaceuticals.

Complement The complement system consists of a series of serum proteins that, for example, are activated by an antigen–antibody reaction via IgM and IgG, with a resulting cascade-like activation, at the end of which a protein complex is formed that can dissolve cell membranes and thus destroy bacteria.

5.4 Complement Fixation

After complexing to a corresponding antigen in the organism, antibodies can bind specifically and so activate a series of serum proteins, which are known collectively as the "complement

system". This activation leads to a mainly enzymatic amplification cascade with the formation of a terminal cytolytic protein complex, which can lyse cell membranes in the vicinity of the antibody binding. It can lyse its own nucleated eukaryotic body cells (cytolysis) and erythrocytes (hemolysis) but mainly – and this is its main function for the protection of the organism – microorganisms (bacteriolysis). In this defense system the antibody system has the function of antigen recognition and the complement system has the function of cell membrane lysis and thus cell destruction. Nevertheless, only certain antibody isotypes activate the complement system (in humans these are IgM, IgG1, and IgG3), which is especially taken into account in the production and usage of monoclonal, humanized, and fully synthetic antibodies (Figure 5.27). Similar differences in complement activation are also found in immunoglobulin isotypes of individual animal species. Furthermore, the complement system can cross the species barrier in some cases. The most sensitive complement activator is IgM, which (as pentameric molecule) already possess vicinal Fc-fragments that are necessary for complement activation (after antigen binding). The Fc-density necessary for complement activation (binding of the complement component C1q) by IgG antibody is first reached with attainment of a corresponding density of antigen. Since complement components bind stoichiometrically to antigen–antibody complexes, one can use them for characterizing and quantifying both antigen and antibody, especially because of the extraordinary sensitivity.

Complement fixation – as this reaction is named – can be set up to quantify one of the three partners (antigen, antibody, or complement) if two of the partners are quantitatively defined (Jules Bordet, Nobel Prize in Physiology or Medicine 1919). The most commonly set up system is hemolysis, using red blood cells. Here, one distinguishes *direct* complement lysis with the use of antibodies against erythrocyte surface antigens from *indirect* complement lysis with antibodies directed against artificially bound antigens on erythrocytes. In this method, the concentration of released blood coloring (hemoglobin) is the value measured for the strength of lysis. For example, the titer of an antiserum can be determined by means of the lysis of half of the test erythrocytes (erythrocytes with complement activating antibodies, against surface antigens).

Complement fixation plays an important role in medicine for the discovery of infectious agents, but it has been (despite its high sensitivity) displaced by radioimmunoassay (Figure 5.17) and this, in its turn, has been mainly displaced by immune assays (Figure 5.24). Since complement is initiated *not only* by antigen–antibody complexes (classical activation) but also through the alternative activation path, which can be triggered by certain bacteria (lipoprotein, protein A), yeast substances, polyanions, and by some viruses, it can be understood that this test is extremely subject to disturbance and therefore considerably more tedious technically than the other two methods just described.

Complement fixation is therefore carried out in glass reagent vessels that must always be freed from disturbing influences by chromic acid pretreatment, which means that considerable effort is involved to remove lipoprotein and bacteria as a requirement for consistent analytical results. For the isolation of certain cells from mixed cell populations, complement-induced lysis with complement-binding monoclonal antibodies for elimination of certain unwanted cells has remained an established standard method to the present day.

5.5 Methods in Cellular Immunology

This section is concerned with the identification and isolation of certain cell types by means of immunological techniques. As in the case of soluble antigens, intracytoplasmic and cell surface antigens can also be employed as labels for the identification, closer characterization, quantification, and isolation of cells with somewhat differing developmental stages or functional states. These labels (or label types), which indicate an increasing complexity ranging from bacteria up to higher multicellular organisms, can today be reliably identified by means of an abundance of commercially available monoclonal antibodies and can be followed throughout the development of the cell system. In this way, the differentiation of complete cell pedigrees can be followed by immunological techniques, starting from the zygote up to the fully differentiated organism, that is, from its origin to attainment of full age. The development of this analysis possibility has expanded continuously and is not yet complete. In similar ways (and additionally with the help of sequencing techniques) the sum of all proteins (proteomics) and

similarly the corresponding genes (genomics) can be described. The final goal is to identify within a cell all genes and their products and all cooperating protein families (functional cluster), and to analyze their individual functions and their successive interactions in order to learn their roles in the process of life. The prerequisites for the attainment of this goal are efficient techniques for cell separation.

Cell isolation should, if possible, be done without artificial stress or cell stimulation. Lymphocytes from the blood of humans and animals can be separated by means of density gradient centrifugation. The necessary density is provided by means of the density medium *Ficoll Isopaque*, so that the less dense lymphocytes remain above the density medium, whereas the heavier erythrocytes and granulocytes are centrifuged through the density medium to the bottom of the tube and, hence, are separated from the lymphocytes. Cells can also be separated with the help of immune adhesion. Cells labeled specifically with an antibody can adhere to a plate on which an anti-antibody (secondary antibody) is bound. The non-adhering cells are then separated by rinsing away, a method known as panning. One can also bind such cells via antibodies absorbed on nylon wool in an affinity column and achieve unhindered passage of non-labeled cells. The secondary antibody can, however, also be bound to an iron-containing carrier molecule that selectively separates the labeled cells from a heterogeneous cell population via the antibody bridge, in conjunction with a magnet (indirect magnet selection). The magnetic particles can also be loaded with the first antibody (direct magnet selection). Human T-cells can also be separated by means of sheep erythrocytes, against which the T-cells carry receptors. The sheep erythrocytes surrounding the T-cells (direct rosetting) alter the density of the complex so that they can be separated from other cells by means of a density gradient. Erythrocytes can also be loaded with antibodies and so allow the technique of inverse rosetting. Other methods eliminate the unwanted cells by means of complement fixing antibodies and complement-mediated cytotoxicity (see above) or via immunotoxins (with toxin-loaded antibodies).

The most elegant method by which to both analyze complete cell populations in unsegregated condition and isolate the desired cells (i.e., *cell recognition* and *cell separation in one step*) is with a fluorescent activated cell sorter (FACS). This complicated apparatus (Figure 5.26a) combines four essential functions. Firstly it consists of a vibration chamber that generates from a cell suspension a series of fine droplets that simply fall under gravity into a collection vessel if they are only to be measured and evaluated statistically. These droplets are so small that they offer, at the most, room for only one cell. In this way, practically all cells are isolated and are thus available for subsequent single-cell analysis and selection.

The apparatus consists secondly of a measuring unit that can determine cell size from the forward scatter (FSC) of a 488 nm (blue) laser beam and determine cell granularity from its side scatter (SSC), measured at 90° angle to the beam (not depicted in Figure 5.26a). Cell surface labels are presented to the sorter by means of fluorochrome-labeled antibodies (this function is not shown in Figure 5.26a). Certain functionally-different cell types can be separately analyzed and isolated in a second dimension by means of labeled antibodies according to their surface label (CD-marker).

The FACS apparatus consists, thirdly, of a collection unit that is attached to the measuring unit. According to the measured value, the corresponding cell receives either a positive or a negative charge or otherwise no charge at all; depending on this treatment the cell is deflected by subsequently-placed charged deflection plates into the appropriate collection tube if it experiences a deflection impulse. In this way, one can isolate up to 10 000 cells per minute from a complex mixture for further analysis or functional testing.

The fourth unit is represented by the electronic computer with its program, which controls the apparatus and converts the surfeit of measurement values into usable data. An example is shown in Figure 5.26b; the stimulation inhibition of granulocytes by serum amyloid A (SAA). The stimulant used is the bacterial tripeptide N-formyl-Met-Leu-Phe (FMLP). The ordinate shows the FMLP-induced oxidative burst, as measured by the fluorescence (FL1) from dihydro-rhodamine (DHR). The abscissa shows the separation by means of SSC (side-scatter) and measures each cell's granularity. FACS sorting results in the separation of granulocytes into three rectangular fields: lymphocytes (4), monocytes (3), and granulocytes (2).

This technique of electronic single-cell analysis of large cell populations and the evaluation of their functions and changing functional status (especially, now, with the additional possibility of measuring a large number of intracellular functions and their pharmacological influence) has greatly widened our knowledge of the dynamics of multicellular organisms, from their origin as zygotes through their development into old age.

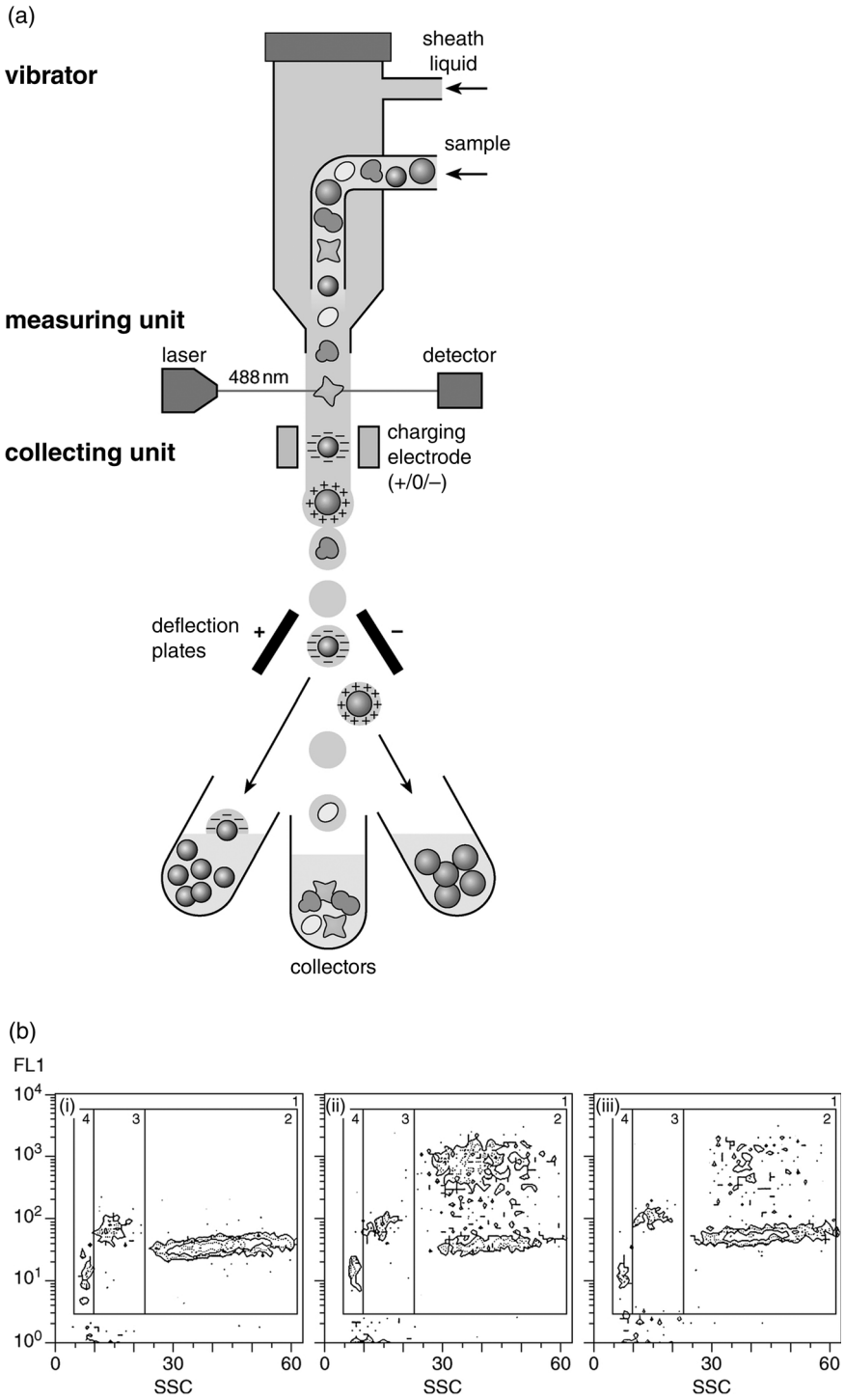


Figure 5.26 FACS, fluorescence-activated cell sorter, schematic. The FACS apparatus can identify and isolate cell populations. (a) Principle of cell separation: The vibrator generates fine droplets that contain one cell at most from a suspension of single cells (sample and sheath liquid). The separated cells are identified by means of the measurement unit, each charged according to their measured values and sent with the deflection plates into the corresponding collection vessels. (b) Example of a functional analysis of unseparated white blood cells by means of the cell sorter (FACS). Shown is the inhibition of stimulation of granulocytes by SAA (Serum amyloid A). The stimulant is FMLP (a bacterial peptide). The oxidative burst was measured by fluorescence (FL1) emitted from DHR (dihydro-rhodamine). Abscissa: separation according to SSC values (side scattering, which measures cell granularity) allows the separation of lymphocytes (4), monocytes (3), and granulocytes (2). (i) Start conditions, non-stimulated cells: the monocytes show a base activity that is higher than that of the granulocytes; (ii) mild stimulation shows a response only from the receptive granulocytes, while the other cells are not stimulated; (iii) in the presence of SAA the FMLP-induced stimulus is largely absent.

5.6 Alteration of Biological Functions

Antibodies can be used for the modification of numerous biological functions, such as for the elimination of toxin effects (toxin neutralization), of pharmaceuticals in overdose, of enzymes, and also of viruses and bacteria. Antibody-based immunopharmaceuticals against bacteria will probably experience a renaissance in the future, since the antibiotics presently still effective are increasingly losing their effectiveness, owing to increasing development of resistance. This is

possibly the result of their misapplication in medical practice and in the veterinary and agricultural domains. Antibodies will also therefore present a second therapeutic option, since the industry can hardly keep up with the preparation of new antibiotics owing to the extremely high costs. Consequently, the industrial production of antibodies against pathogens could become economically interesting, the more sophisticated and simpler the techniques of immunopharmaceuticals become. Such products are now in development and have already recorded quite decisive successes (Section 5.7.2).

These antibodies cooperate together with the effector mechanism already present in the organism (Fc-receptor functions such as complex-formation and phagocytosis, complement-lysis, antibody-mediated cytotoxicity). There are antibodies that activate inactive enzymes, also some that can inactivate active enzymes. These functional alterations can also be employed for planning biochemical tests (Figure 5.24). Antibody-mediated options are mainly used in cellular immunology and cell biology to influence therapeutically a great number of diseases. In this respect, *modern antibody technology* has developed, by means of protein-chemical and gene-technical methods, to produce highly efficient and compatible immune pharmaceuticals of a quality that seemed unthinkable only some decades ago (Section 5.7.2).

5.7 Production of Antibodies

Although today a large number of excellent isolated immune sera and multiply-labeled antibodies are commercially available, there are of course still no antibodies for the latest discovered proteins or for proteins of still insufficient commercial importance. Therefore, presented here is the simple possibility of self-produced immune reagents. This requires at the outset access to an officially registered animal laboratory, which is attended by a veterinarian, and/or to a tissue culture laboratory for the production of monoclonal antibodies, if the antibody or the Fv-fragments (V-modules) are not gene-technically produced. Antibodies can be produced against proteins, polysaccharides, glycolipids, and almost all chemical groups, which, however, must satisfy several requirements to be immunogens (Box 5.2). This section will be limited to the production of proteins and haptens, since these represent at present the most important immunogens.

5.7.1 Types of Antibodies

To be able to produce antisera, quite differing laboratory animals are immunized (mouse, rat, guinea-pig) as well as domestic animals (rabbit, goat, sheep, pig, horse, lama, hen; listed here more or less in order of importance). These animals are specially bred and offered commercially for yielding antiserum. The antigen is prepared and applied for immunization in various ways. Preparation includes the use of adjuvants to boost the immune response and thus increase the antibody concentration. The first application (immunization) is mostly done strictly subcutaneous, at multiple sites in the form of micro deposits, which hardly affects the experimental animal. The resulting antisera contain a large amount of various antibodies with different epitope specificities and affinities, whose composition changes according to blood collection time and the number of *booster* injections (subcutaneous or intramuscular) that take place approximately monthly after the first immunization and which also contain adjuvant through antigen-induced expansion of ever-newer plasma cell clones.

Antisera contain polyclonal antibodies, which react with the homologous antigen, as depicted in the models of Figure 5.7a and 5.7c. The only differences are that a natural antigen usually has more than three antigen determinants, and that a polyclonal antiserum can, as a rule, react with a battery of about 10–1000 various antibodies, each against single antigen determinants, whose number and individual synthesis performance can change after each booster injection, depending on the individual animal and its genetic composition.

Thus, an antiserum can never be exactly reproduced. On the other hand, the antiserum recognizes the antigen in almost all conformations, even after strong denaturation (reductive splitting of disulfide bridges and also in the presence of certain detergents) or after fragmentation, since in many cases enough antigen determinants remain for binding. Antisera are therefore ideal for use in immunoprecipitation methods.

Adjuvant A substance that strengthens the immune reaction and the *antibody titer*. In the case of preventative immunization in humans the immunogens are adsorbed (e.g., to aluminium compounds). For immunization of experimental animals for antiserum extraction, mineral oil, which is injected as an emulsion with the immunogen solution, has proved to be appropriate (incomplete Freund's adjuvant). The addition of capsule substances from mycobacteria further increases the immune reaction (complete Freund's adjuvant; but first consult the animal protection laws concerning this application!). These days only certain subcomponents of the bacterial membrane are used as an adjuvant in order to minimize the inflammation reaction.

Peptide antibodies are polyclonal antibodies against custom-made synthetic peptides of various lengths; they are directed against well-defined, mostly functional or structurally-important regions of a macromolecule, such as for the investigation of catalytic or other binding functions. The reaction of peptide antibodies corresponds to that depicted in Figure 5.7b, except that these react against a single or few epitopes, but with a large number of different antibodies. Peptide antibodies are also produced pair-wise for bimodal binding of a molecule in certain assays (e.g., in sandwich-ELISA, Figure 5.24).

Owing to its monovalent binding to antigens (Figure 5.7b) a peptide antiserum does not precipitate, but it is highly specific in binding tests, provided that the peptide epitope is the immunogen homologue or chemically very similar.

Peptide antibodies are produced after coupling peptides to immunogens, multimolar carrier molecules like thyroglobulin or KLH (keyhole limpet hemocyanin, *Megathura crenulata*, or the hemocyanin of the Atlantic horseshoe crab *Limulus polyphemus*).

Hapten antibodies, which can be produced in a similar way against smaller chemical substances (like steroid hormones, pharmaceuticals, drugs, environmental poisons), generally bind only with a part-segment of the paratope (paratope subsite) (Figure 5.2c).

Monoclonal Antibodies An individual B-cell produces only a single constant antibody specificity in the organism (Niels Jerne, Nobel Prize in Physiology or Medicine 1984), which does not alter, even with isotype change. If a B-cell degenerates neoplastically, resulting in unlimited proliferation, an exorbitant amount of an antibody can be produced (homogenous antibody). The disease responsible (e.g., multiple myeloma, plasmacytoma) can be diagnosed in the presence of this one immunoglobulin in the blood plasma (Figure 5.13). Here, it usually concerns a single normal antibody in excessive amounts, whose specificity is known only in exceptional cases.

In 1975 George Köhler and César Milstein (Nobel Prize in Physiology or Medicine 1984) were able to produce in the mouse, by means of somatic hybridization (fusion) of immune-B and tumor cells, monoclonal antibodies of intended specificity with unlimited growth duration in tissue culture. Using a chemical selection mechanism allowing only growth of fusioned cells (hybridoma) and functional selection by antigen binding, only clones with the desired antibody specificity were propagated.

The hybridoma technique has revolutionized immune detection, since now antibodies of chemically identical structure and function are available in potentially unlimited amounts. These antibodies are used in manifold immune assays and are ideal reagents for automation. In addition, the production of monoclonal antibodies in tissue culture assists our goal of minimizing experimentation on animals. Because of these especially favorable characteristics, monoclonal antibodies have won growing importance in all biosciences, chemistry, medicine, and other areas of the industry.

Isotype change After first contact of an organism with a foreign antigen, IgM-antibodies are generally formed (Section 5.1.3). With further antigen exposure, additional IgG antibodies and antibodies of other immunoglobulin classes are produced. Thus, a change of immunoglobulin class occurs that is termed an isotype change (*class switch*). Here, only the constant part of the immunoglobulin is exchanged, that is, the variable section and the light chain of the immunoglobulin molecule remain unchanged (Figure 5.1), as also does the antibody specificity.

Monoclonal antibody Antibody of chemically individual structure and function, produced *in vitro* from a cell clone created through fusion of immune plasma cells with induced specificity (of limited life span) and potentially eternally-living tumor-B-cells (Section 5.4).

Polyclonal antibody Antibody against an antigen, which is formed from many cell clones and which represents many varied isotypes with varied affinity to different epitopes. Antibodies induced through immunization are mostly polyclonal.

B-Cells B-cells are certain lymphocytes that differentiate to antibody-producing cells (i.e., B-lymphocytes and plasma cells). These cells generate *humoral immunity*. By contrast, T-cells confer *cellular immunity*.

5.7.2 New Antibody Techniques (*Antibody Engineering*)

The starting point of the new antibody techniques is monoclonal antibodies, since they are chemically identical. With the production of monoclonal antibodies the use of analytical and preparative antibodies in chemistry, biosciences, and medicine took a huge upturn, because through the possibility of routine production of chemically uniform reagents against practically any given antigen in unlimited amounts a hitherto unknown level of precision and an astounding range of applications could be attained. By using this technique, the generation of antibodies has come very close to that of a chemical reagent with consistent structure and function. Therefore, as well as highly precise and automatic *in vitro* test systems, new *in vivo* diagnostics and *in vivo* therapeutics became possible. These antibodies came close to the dream of Paul Ehrlich (Nobel Prize in Physiology or Medicine 1908) of a highly specific therapeutic reagent (the "magic bullet").

While *in vitro* test systems with monoclonal antibodies (mostly from the mouse) experienced a fast and steep upswing, the first successes with *in vivo* applications on humans came first after a delay of one or two decades. This was because even a uniform antibody molecule always represents a *multifunctional* as well as *species-specific* protein that is an immunogen in other species (see below). This multifunctionality is described in Section 5.1.4 and concerns the

idiotypes, allotypes, and isotypes of individual antibodies. A single antibody molecule plays only certain part-functions of the multifunctionality in the overall extraordinarily heterogeneous antibodies in an individual; nevertheless, all functions of an individual antibody in the organism are characterized by varying humoral and cellular mechanisms in functional exchange. Every single antibody during its synthesis in each individual is therefore quality-checked for its compatibility with the complete organism; if a B-cell can produce no compatible antibody for the organism it will become aborted. Therefore, all antibodies present in an organism are regarded as "in harmony with the immune system concert" when they satisfy the very low blood level required of each individual paratope. The way in which this internal quality-control occurs is not yet known. It is, however, a central question in the production of immune pharmaceuticals for use with humans, since this test of compatibility with the individual to undergo therapy currently can first occur only *post festum,* which means after industrial production of the extremely expensive antibody. While one can *in vitro* bring only one or another part function into use or test through extreme simplification of the testing approach, the antibody molecule must be so structurally altered for the application *in vivo* that it alone fulfils the function ascribed to it and executes no other function during its "natural life". Hence it was a requirement to optimize each of these part-functions in isolation, which in essence worked by means of gene-oriented techniques. Here, a few examples are presented, which reveal new developments in the field of antibody development (*engineering*), which after decades have finally led to the first successes. As long as the production of human monoclonal antibodies by the hybridoma technique (as with the mouse) or virus-induced immortality of human B-cells and those from individuals with microbial inflammation remained unsuccessful, murine monoclonal antibodies were used, since their production is today possible in every tissue culture laboratory. *In vivo* application of monoclonal mouse antibodies in humans was prevented by their *immunogenicity,* which via an immune response not only led to fast elimination with loss of effectiveness in patients but also in some cases led itself to patient injury through the development of life-threatening allergic reactions. The immunogenicity problem was countered in two ways, first by means of manipulative humanizing (Figure 5.27e–h) and secondly by reducing the size of the antibody (Figure 5.27b, c, j–l).

Bispecific antibodies, hybrid antibodies Normally an IgG antibody is divalent, but monospecific, since it consists of at least two identical (and not mirror-imaged) monovalent antibodies (Figure 5.1). If, however, an antibody unit combines two antibodies with different antigenic specificities, one has again a divalent, but bispecific, antibody. These antibodies are not of natural origin, but are artificially produced, for example, by releasing the covalent S–S-binding of both H-chains in a mixture of two monoclonals with different specificities through mild reduction and subsequent re-oxidation. In the antibody product there is then found hybrid antibodies as well as the previous monoclonal antibodies (Figure 5.27i; 5.28c(i) and 5.28d(i)).

scFv-Antibodies (***single chain antibodies/fragments***) Single-chained antibodies. These consist of the two Fv domains VH and VL, which are covalently bound together over a peptide bridge containing Gly–Ser (Figures 5.1 and 5.28c(iii)).

Chimeric antibodies Antibodies, which are assembled from immunoglobulin regions of different species (e.g., mouse Fv and human CH1-3 (Figure 5.27f) or bispecific antibodies (rat-mouse, Figure 5.28d(i)).

Recombinant antibody Gene-technically produced antibodies that can represent classical antibody fragments, *Mini-antibody,* or *continuous antibodies.*

With the stepwise replacement of murine antibody regions with corresponding regions of a human immunoglobulin by means of protein-chemical and gene-technical methods, the antibody became (through the intermediary step of chimeric antibody) more and more compatible, until the immune response was so far reduced that ever fewer humans reacted against repeated injections with an incompatible immune response. At present an average reactivity of 3–20% is given, which has, however, with humanized antibodies fewer consequences than with chimeric antibodies since they no longer lead to rapid elimination with complete loss of effectiveness, but (perhaps mainly by the existence of small soluble immune complexes, say against paratope-external idiotypical antigen determinants) do not essentially reduce their antibody effectiveness.

The first successful therapeutically-applied antibodies were chimeric antibodies, in which only the Fv-part (Figures 5.1 and 5.27) remained from the mouse, and the rest, the strongly immunogenic two-thirds of the antibody (C_1–C_3, in Figure 5.1; Figure 5.27f), was of human origin.

To humanize the remaining murine third of the Fab-region only the antigen-binding paratope of the mouse was taken, which led to the so-called humanized antibody with about 90% human content (Figure 5.27g).

These antibody constructs have led to successful therapeutic antibodies, even when immune reactions are still present. A further type of humanization was achieved by gene-technical replacement of all murine parts by human parts. This method, which consists of a point-by-point humanizing (carried out *in silico* and called resurfacing) so that the paratope is also humanized, has led to a completely *humanized antibody* (Figure 5.27h), which nevertheless is reconstructed from a mouse prototype. To circumvent the very laborious indirect route involving the mouse, a way has been found to produce a recombinant mouse, which alone expresses the complete immune system of humans, so that it is possible to acquire human antibodies directly in this mouse. Another possibility has also been achieved, namely, the growth of single human B-cells from patients with certain infectious diseases and tumors. In the last case one can already achieve protective antibodies, which must only be expanded if allotypes are not disturbing.

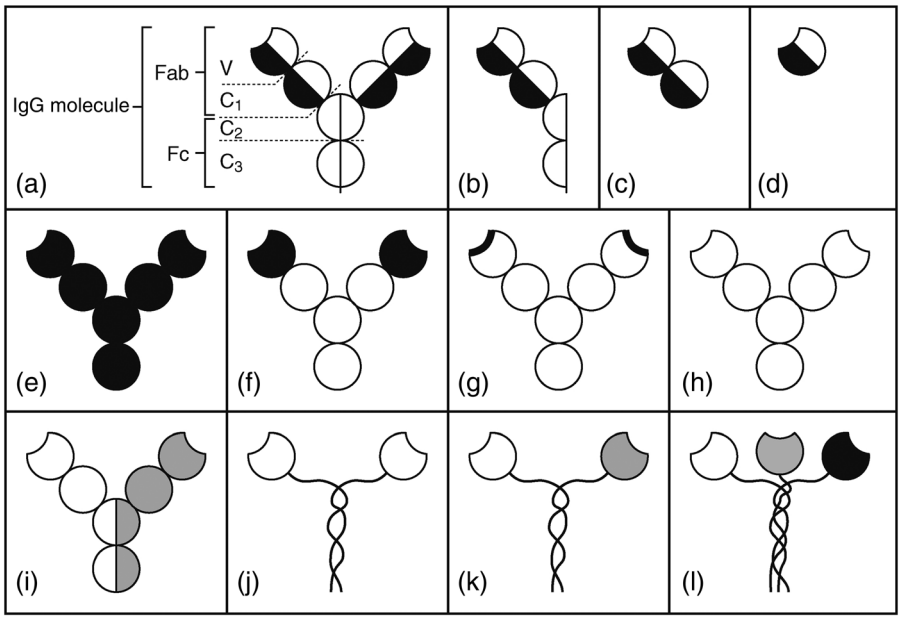


Figure 5.27 Antibody-engineering. (a)–(d) Black: L-chain; white: H-chain. (a) Antibody molecule, very schematic (compare Figure 5.1); (b) monovalent antibody-half; (c) Fab-fragment; (d) Fv-fragment, two-chained; (e)–(h) step-wise humanizing of a mouse antibody; black: murine, white: human; (e) pure mouse antibody; (f) chimeric antibody; (g) humanized antibody; (h) fully humanized antibody, natural human antibody or fully synthetic human-antibody; (i)–(l) white: specificity 1; grey: specificity 2; black: specificity 3; twisted lines, hydrophobic polypeptide; (i) bivalent bispecific hybrid-antibody; (j)–(l) mini-antibodies: (j) monospecific, divalent; (k) bispecific, bivalent (hybrid-mini-antibody); (l) tri-specific, trivalent (hybrid-mini-antibody).

Usable results have been achieved with both these methods. They are nevertheless still a limitation, in that only antibodies can be prepared against substances which can be recognized by mouse or human as immunogens, and indeed also because of the limitation of the gene pool in just one animal (or a limited number of animals), which can never represent the range of variants of the total species' genome. It should be added that because of the limiting stochastics of somatic rearrangement and affinity maturation by somatic recombination and point mutations the striven-for specificities themselves do not always attain the expected quality (epitope faithfulness and highest affinity). Therefore, new directions have been taken to produce human antibodies without the use of immunization. A way favored at present is the gene-technical production of fully synthetic human antibodies. To ensure that these antibodies are tolerant for humans the total human gene pool in gene-banks was made available from which complete (and compatible for the desired organism) gene modules are on call. The starting material for the paratopes are proven gene-banks, from which the required specificity is selected by competitive mechanisms from randomly-generated variants. They therefore are no longer subject to the restrictions of V-genes of single individuals. High affinity is thus achieved by the selection of corresponding variants (affinity maturation *in vitro*).

By means of available modules that represent all γ-isotypes and their *framework*-variants, as well as the large number of complementarity determining regions (CDRs), we have a set of "building blocks" from which any desired antibody can be assembled and then can be "tailor-made" via fine selection (epitope stability, affinity, format to something like scFv-antibody or Fc-*engineering*) to suit the functional requirements.

The three direct ways to therapeutic human antibodies (immunopharmaceuticals), obtained by evading immunization of the mouse, have today already led to most therapeutic antibodies that are on the way to becoming possible therapeutics. These ways are thus determining the future. Which of the various ways (humanizing of mouse monoclonals, mouse with human immune system, immortalized human B-cells of immune individuals and fully synthetic antibodies, also gene banks derived from the mouse with the help of human gene banks) finally delivers the best immunopharmaceutical for a particular application is at present still unknown. Since the branch of industry dedicated to the production of therapeutic antibodies already serves a billion dollar market, which can boast a multi-digit yearly growth rate, practicality is measured in terms of production speed and the price/performance ratio.

To establish an antibody as a standardized immunopharmaceutical its *in vivo* application has to be tested experimentally after its production (the reasons for this are given above). Then, even a completely human antibody must itself be subject to *regulation by the individual organism.* This fact has the effect that an antibody is only available in a restricted compartment in the organism,

which can prevent the attainment of the therapeutic goal. In addition, countermanding by the organism of an antibody-related functional alteration can prevent the desired effect of the antibody. Furthermore, the organism responds to every idiotype with higher concentration (mono-paratype) of a corresponding anti-idiotype, even with the organisms' own self-synthesized antibodies (AK2-antibodies of Jerne's "network theory"). An example is the disease-complex of monoclonal gammopathies, in which higher concentrations of AK2-antibodies are found, but not to the same extent in each individual. When it happens that the self-produced antibodies of an organism can invoke such an immune response, then an immune response to fully synthetic human antibodies must not necessarily be the result of inadequate preparation technique since all antibodies in a network are regulated and the organism's monoclonality, which means mono-paratype with subsequent auto-reactivation against the paratope (with loss of effectiveness) is effectively avoided, which may contribute to the retention of as many specificities as possible. It is therefore important to verify that up to now the immunogenicity of a synthetic or natural monoclonal antibody is unpredictable and therefore (as with all first applications on humans) must in every case be tested *in vivo* before therapeutic usage. The complete testing with respect to compatibility and therapeutic effectiveness is done in the clinic in different stages, which are named preliminary experimentation, phase I, phase II, and phase III.

A further possibility to improve compatibility is to reduce the size of the antibody. This allows optimization also of part-functions. In this way antibody functions can be mediated through Fab fragments or Fv fragments (Figure 5.1, 5.27b–d and Figure 5.28c(iii)). An Fv-module can consist either of two independent domains (VL + VH) or of covalently-bound Fv-domains in the form of a polypeptide chain (scFv, *single chain* Fv). These very small antibodies (mini-antibodies) have the advantage of penetration into body regions that are unreachable with intact antibodies, and also the advantage of lacking antibody sections that are associated with many *in vivo* functions. Another advantage is the possibility of multimerization of a single antibody function with in turn one or more Fv-modules of the same specificity (divalent and monospecific) by means of hydrophobic tails (Figure 5.27j–l). Moreover, it allows the choice of any combination that is not foreseen *in vivo*, for example, the combination of two, three, or more specificities (bi-, tri-, and multispecificities). The disadvantage of mini-antibodies is their short half-life since they can easily diffuse into all (also therapeutically irrelevant) regions of the body and also especially when they can be quickly eliminated via the kidneys.

The successful production of a therapeutic antibody requires firstly an antigen target that substantially weakens or destroys the tumor vitality and which, for example, acts only on the tumor via vital intracellular nuclear functions and acts less on normal cells. For this, the antibody requires the highest possible specificity and a high affinity; the latter can be attained by affinity maturation *in vitro*.

To overcome mono-paratype, which can lead to reduction of effectiveness via an anti-idiotype reaction (especially in the case of completely human antibodies), one can, according to the organism that produces a complete series of antibodies to avoid mono-paratype and prevent AK2-antibody induction, offer a series of mono-epitope-specific antibodies on the basis of different CDR-regions and different isotypes, as already tested. The removal of an individual immunogenic antibody from a larger mixture of recombinant antibodies can thus lead to an individual compatible cocktail of oligoclonal immunopharmaceuticals, from which no one idiotype stands out quantitatively as isolated. An advantage of these cocktails would also be that various Fcγ-isotypes could be used for the complete restoring of the usual multiple Fcγ-effector functions so that the antibody effect can be increased further.

A still speculative idea is the possibility of the phenomenon of an "accustoming" of an organism to the application of a mono-paratope via the development of an anti-idiotype, which should not disrupt much the effectiveness of these immunopharmaceuticals. Since the bodily anti-idiotypic antibodies carry the internal structure of the target epitope, it is possible that the organism starts up a polyclonal response against these anti-idiotypes and thus against the target, which could explain the phenomenon of therapeutic permanence.

5.7.3 Optimized Monoclonal Antibody Constructs with Effector Functions for Therapeutic Application

This section presents some typical examples with optimized IgG antibody constructs, which are already in clinical use, or on their way to being used (Figure 5.28). As well as the

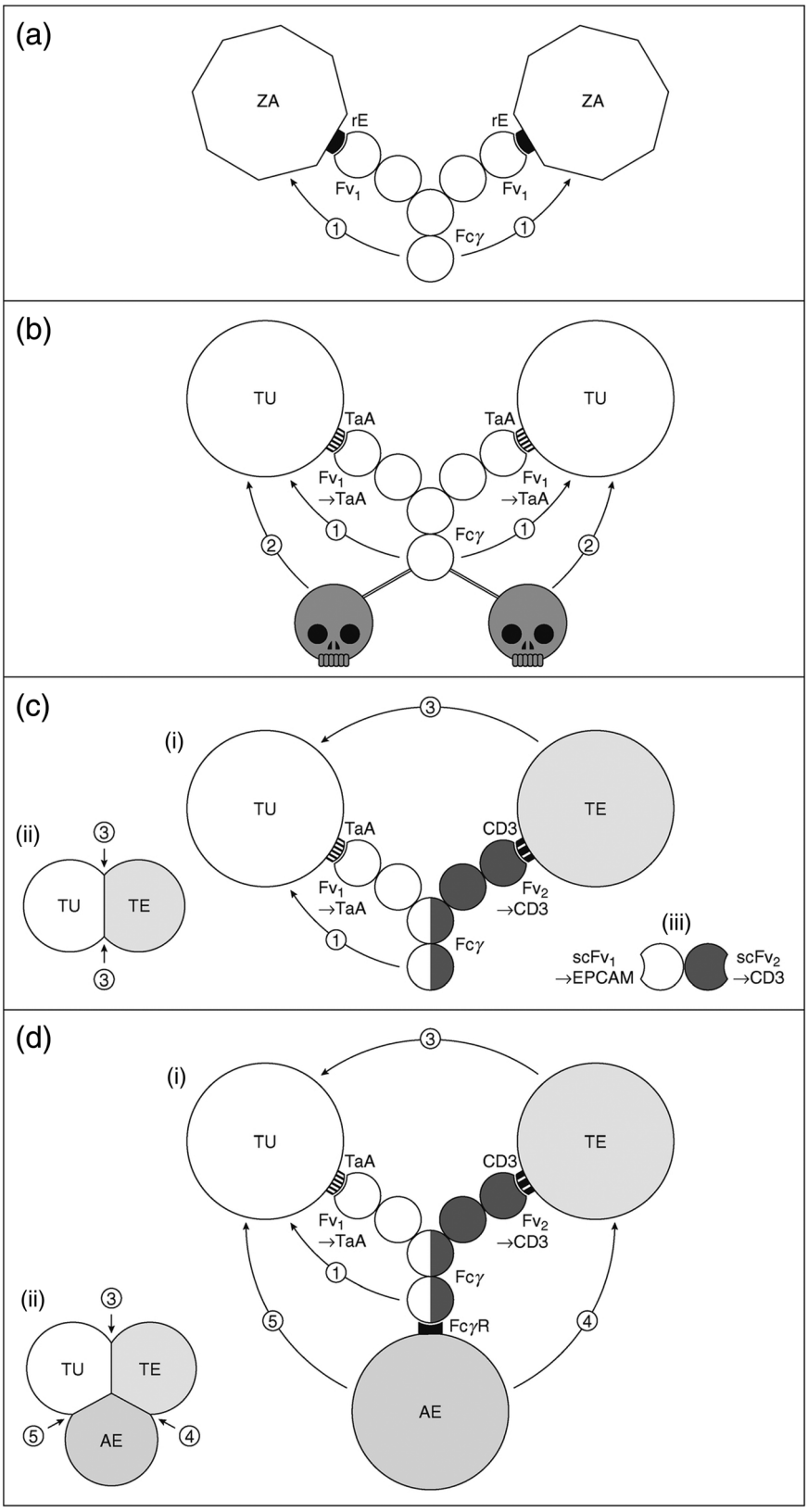


Figure 5.28 Four known strategies for the production of immunopharmaceuticals on the basis of gene-technical production of monoclonal antibodies. (a) Imitation natural antibodies with natural effector functions; (b) antibodies as in (a), with artificial covalently-bound toxic substances; (c) recruiting of cytolytic effector cells by means of bispecific antibodies; (d) trifunctional antibody with recruiting of cytolytic effector cells as in (b) and with recruiting of Fcγ-associated effector cells. Abbreviations: ZA, target antigen; rE, relevant epitope; TU, tumor cell; TE, T-effector cell; AE, associated effector cell. Functions (see circular arrows): 1: Fcγ-functions; 2: artificial cytotoxic functions through "poisons"; 3: cytolytic T-cell functions; 4: additional up-regulation of TE; 5: AE-assisted TU-Lysis.

natural effector functions of intact antibodies there are also artificial super-cytotoxic functions ("poison" in its most general meaning), which strengthen the effector functions. At present the most effective seems to be the recruitment of natural killer cells, whose cytotoxic principle is about a thousand-times more efficient than the effector functions of intact antibodies. Two systems in the third and fourth examples are cited. These systems lyse very efficiently loose tumor tissue and detached neoplastic cells in blood and ascites. The problem of non-optimal penetration of intact antibodies deep into solid tumors is being tested by employment of mini-antibodies. Examples are already in clinical trials (Figure 5.28c(iii)).

Antibodies in Conjunction with the use of Natural Fcγ-Effector Functions Figure 5.28a shows divalent monospecific antibody constructs without additions (naked antibody of 150 kDa), whose natural effector-functions (Fcγ-transmitted, Figure 5.1) can be used for therapeutic purposes. Amongst these are the target antigens (ZA, Octagon), almost all natural antigens (like those on bacteria), viruses, tumor viruses, fungi, and parasites (which under certain conditions can also be immunized against). Target antigens are also, and especially today, those on neoplasms and the effective proteins of quite various systems (e.g., clotting in heart attack, erythrocyte stability in hemolysis, prevention of the fever-cascade in autoimmune diseases, angiogenesis, and the prevention of tumor growth).

For each of the above systems, the recognition of the functionally dominant and thus relevant epitope (rE) is initially a problem that by means of specific binding via a paratope on the Fv-section (Figure 5.1) can identify the tumor type. According to the binding-initiated Fcγ-induced effector functions (1) are complement-induced cytolysis, phagocytosis, and antibody-directed cellular cytotoxicity (ADCC). With the use of additional effector functions (Figure 5.28b–d), these natural Fcγ-mediated effector functions remain effective, without that they are contributing artificial effector functions (Figure 5.28b) or the additional cytocidal functions (Figure 5.28c,d). The natural antibody functions described here require the intact antibody, which cannot be made smaller.

Immune Conjugates These antibody constructs correspond to those discussed previously, but with an additional artificial effector function, which here is simply summarized as "poison" (death skull in Figure 5.28b). These immunoconjugates can multiply increase the natural effector functions through the covalently-attached "low-molecular-weight poisons" (<1000 kDa). In this method the poison is either bound over synthetic linkers to random sites of IgG (about three per IgG) or the placing of the synthetic linker is quite deliberate, to protect the functions of the CDR region (antibody binding function) and that of the Fcγ-region (effector functions). The placing (site-directed binding) can be achieved over the polysaccharide of the C3 module, over the available (or specifically-added) SH-functions, or over specially selected lysine bindings. The linker can be hydrophilic, hydrophobic, stable, acid-stable, or sensitive for lysosomal enzymes, the latter two each according to the selected working mechanism. The effect striven for is the antibody binding on a TaA antigen with coupled internalization and intracellular separation of the poison from its IgG carrier. With this, the poison can bring the cell to apoptosis, either through the destruction of nucleotide functions or through the actin-myosin structure. The poisons used are very abundant and basically are still in the development and testing stage. There are natural poisons such as ricin, vinca-alkaloids, taxoids, and the chemically very toxic cytostatic medicine doxorubicin. The special poisons used for this purpose have nevertheless a toxin power that can lie 2–4 orders of magnitude over that of doxorubicin. This high toxicity is directed not only against tumor cells but also against normal cells (bone marrow, liver) and it is extremely difficult (requiring great artistic skill) to ascertain exactly the therapeutic window needed for success with *specific* tumor destruction. Successes are then to be expected if the TaA can be found in much higher concentration on neoplastic cells than on normal cells. Poison, in the wide sense used here, can also include radionuclides (^{131}iodine, ^{111}indium, ^{90}yttrium), which likewise strengthen the therapeutic efficiency of antibodies, but can also show severe side effects (see above). Quite generally, the production of immunoconjugates demands a greater biochemical and financial expenditure up to the production of individual molecules

in large numbers in unchanging quality – difficulties that can limit their therapeutic application.

Recruitment of Cytolytic T-Cells over Bispecific Antibody Constructs The system depicted in Figure 5.28c consists of two antibody-halves, each with its own antibody specificity. This bivalent antibody is thus bispecific and is a hybrid antibody. The arm with Fv_1 is specific for TaA on the tumor cell (TU), like that already shown in Figure 5.28b. Likewise, the second arm with Fv_2 is specific for CD3. The second arm recruits CD3-T effector cells (TE), resulting in an activation of a TE cell. The cytolytic "hit" is carried out over the close apposition of both cell membranes via function (3) in Figures 5.28c(i) and 5.28c(ii). The killer T-cell remains intact and can thus lyse more tumor cells, resulting in a somewhat prolonged action.

Figure 5.28c(ii) is a schematic representation of antibody-induced pairing of the tumor cell with a cytotoxic T-cell with large area membrane adhesion, resulting in the highly effective *kiss of death* (3). Since this system requires about one thousandth the amount of antibody to achieve a similar effect as the Fcγ-dependent lytic effector functions (Figure 5.28a), the species difference between human and mouse (bi-specific mouse antibody) plays practically no role.

A successful construct in this manner is the BiTE system (Bi for bivalent and TE for T-*engaging*), which has already progressed as far as successful clinical tests. Better penetration in solid tumors is sought using mini-antibodies, starting from two monoclonal mouse antibodies with renunciation of the whole constant parts of both H and L-chains. The mini-format $(scFv)_2$ (sc stands for *single chain*, see above) of about 50 kDa, with antibody properties $scFv_1$-anti-EpCAM and $scFv_2$-anti-CD3, binds to the EpCAM-Antigen on epithelial tumor cells (Figure 5.28c(iii)). Renouncing the Fcγ functions has not abolished the lytic properties of this system *in vitro*. When these mini-antibodies strongly penetrate solid tumors and bind there to the corresponding TaA, they still require the recruitment and activation of cytolytic CD3-T effector cells, which are, however, present in variable numbers.

Activation of Cytolytic T-Effector Cells via the Recruitment of Accessory Effector Cells through Tri-functional Antibody Constructs The system depicted in Figure 5.28d resembles that of Figure 5.28c with regard to the bispecificity of the antibodies and their epitopes. The first specificity is against EpCAM, a tumor cell (TU), and the second is against the CD3 antigen on T-effector cells (TE). The third specificity is facilitated through the Fcγ-part of the antibody, which recruits a human Fcγ-receptor-expressing accessory effector cell (AE). CD3 binding activates the T-cell and Fcγ binding activates die AE cell. The course of activation of AE is the excretion of highly effective immunomodulators (interleukins like IL-1β, IL-6, IL-12, and DCCK1) that (over the activation through CD3 binding) activate and arm the TE beyond (function 4) for the final tumor lysis (function 3). The peculiarity of this hybrid antibody is also its chimerical nature, in that the anti-EpCAM specificity comes from an IgG2a monoclonal mouse antibody and the CD3 specificity stems from a monoclonal rat IgG2b antibody. The Fcγ function is also chimeric, which leads to a most effective recruitment and activation of AE cells via rodent–human cross-reactivity. This leads further to phagocytosis of the lysed tumor cell. The TE not only survives the tumor cell lysis, but can later lyse more tumor cells, as in Figure 5.28c. This high activation leads in some cases even to T-cell propagation with subsequent continuation of lysing effectiveness, resulting in long-lasting immunity. Fcγ R-expressing accessory effector cells represent antigen-presenting cells like monocytes, macrophages, dendritic cells, and antibody-dependent natural killer cells which can support the lytic TE function (function 5). Also important in this system is the quite miniscule amount of antibody, which (as in the BiTE system) with about one thousandth the amount needed to achieve a Fcγ Lysis. One can thereby recognize a high cytocidal potential of this system. The schematic diagram in Figure 5.28d(ii) shows the wide-area contact of the three cells (TU, TE, and AE), which is caused by the binding molecules, similar to the scheme in Figure 5.28c(ii). The Fcγ-Fcγ R association leads to interleukin-amplified accessorized up-regulation (function 4) of the already (over anti-CD3) activated TE, resulting in an intensive lysis (function 3) via the association of EpCAM and CD3. The association of Fcγ and Fcγ R may contribute via accessory Fcγ-functions to the efficiency of lysis and phagocytosis (function 5), with the possibility of a further individual immune response through a presentation of AE-processed tumor antigens via specific T-cells.

5.8 Outlook: Future Expansion of the Binding Concepts

The goal of the systems described is the optimized pharmacokinetics and pharmacodynamics of finalized immunopharmaceuticals, namely, behavior and effectiveness of antibodies in an organism, including the efficiency of effector functions, permanence of effect, half-life, side effects, and immunogenicity. A series of immunopharmaceuticals, based on natural V- and C-modules, have completely fulfilled therapeutic expectations. It remains to be seen which system will continue to be used for future applications in the ever increasing range of indications.

In the meantime, the artificial immune mimics that have been briefly discussed here are being expanded from a *great number of other systems,* including the establishment of alternative gene banks, that are all derived from typical biological binding-proteins. These systems function quite differently from the V- and C-modules discussed here, but they arrive at similar results in terms of specificity, affinity, and stability. These proteins are mostly smaller than immunoglobulins and less multifunctional. Already, small pharmaceuticals have been developed from these new binding variants and await clinical testing. In addition here, as previously, the laws of immunogenicity require obedience in the application on humans because immunocompatibility occupies, as shown, only a "small window".

In this connection, new perspectives for the discovery of previously unknown binding proteins (with previously unthought of possibilities) are also opening up for bioanalysis. Essential insights here would be obtained through the recognition of immunoanalysis and the production of fully synthetic immunoglobulins and will be carried forward and expanded over the new biomimicry of other systems. One can, however, also think of binding systems, which go beyond biomimicry. These can rest on any structure, even beyond that of the proteins. Here are, especially for new generations of scientists, options that will open quite new perspectives, including for bioanalysis.

Dedication

This chapter is dedicated to Professor Dr. Konrad Beyreuther, Heidelberg, and Professor Dr. Robert Huber, Martinsried Germany.

Further Reading

Beck, A., Wurch, T., Bailly, C., and Corvaia N. (2010) Strategies and challenges for the next generation of therapeutic antibodies. *Nat. Rev. Immunol.*, **10**, 345–352.

Frenecik, M. (1993) *Handbook of Immunochemistry*, Chapman & Hall, London.

Garvey, J.S., Cremer, N.E., and Sussdorf, D. (1977) *Methods in Immunology*, W.A. Benjamin Inc.

Hudson, L. and Hay, F.C. (1989) *Practical Immunology*, 3rd edn. Blackwell Scientific Publications, Oxford.

Janeway, C.A., Travers, P., Walport, M., and Shlomchik, M. (2002) *Immunologie*, 5th edn. Spektrum Akademischer, Verlag, Heidelberg.

Klein, J. (1991) *Immunologie*, VCH-Verlagsgesellschaft, Weinheim.

Moroney, S. and Plückthun, A. (2005) Modern antibody technology: the impact on drug development, in *Modern Biopharmaceuticals* (eds J. Knäblein and R. H. Müller), Wiley-VCH Verlag GmbH, Weinheim.

Plückthun, A. (2009) Alternative scaffolds: expanding the options of antibodies, in *Recombinant Antibodies for Immunotherapy* (ed M. Little), Cambridge University Press, New York, pp. 244–271.

Tijssen, P. (1985) Practice and theory of enzyme immunoassays, in *Laboratory Techniques in Biochemistry and Molecular Biology* (eds R.H. Burden and P. H. Knippenberg), vol. 15, Elsevier Science Publishers B. V., Amsterdam.

Weir, D.M., Herzenberg, L.A., Blackwell, C., and Herzenberg, L.A. (eds) (1986) *Handbook of Experimental Immunology, Volume 1: Immunochemistry*, Blackwell Scientific Publications, Oxford.

6 Chemical Modification of Proteins and Protein Complexes

Ferdinand Hucho[1] and Viktor Tsetlin[2]

[1]Freie Universität Berlin, Institut für Chemie/Biochemie, Thielalee 63, 14195 Berlin, Germany
[2]Russian Academy of Sciences, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Miklukho-Maklaya str. 16/10, 117871 Moscow, Russia

Chemical modification of proteins was – and still is – important in protein research. Years ago, during the times of classical enzymology, it was virtually the only means for the investigation of structure–activity investigations. This represented primarily the elucidation of functional groups essential for the catalytic activity. The aim of chemical modification technologies was to elucidate amino acid side chains (e.g., carboxyl groups of aspartate or glutamate residues, imidazole groups of histidine residues, hydroxyl groups of serine or threonine residues, etc.), the chemical modification of which affects the enzymatic activity (reducing or completely blocking it). Subsequently, the position of this so-called "essential" residue within the polypeptide chain was determined.

The accessibility or non-accessibility, respectively, of these functional groups for the modifying agents allowed conclusions to be made as to their position at the surface or the interior of the protein. Later this helped to find the active site within the three-dimensional structure as analyzed by X-ray crystallography. The methods of protein modification and X-ray crystallography emerged in parallel, and in general the results of the latter were in good agreement with those of the former.

NMR Spectroscopy, Section 18.1

Assignment of peaks of NMR spectra, originally a classical application of chemical protein modification, is now accomplished more efficiently by site-directed mutagenesis. Furthermore, the rapid progress in NMR analysis allows today complete assignment of signals and the elucidation of ever larger protein structures by NMR techniques directly without chemical modification and site-directed mutagenesis, respectively. In complicated situations, for example, when site-directed mutagenesis causes significant structural changes it might still be necessary to resort to chemical modification. But more important for NMR structure analysis is, nowadays, labeling of proteins with ^{15}N- and/or ^{13}C-atoms by protein biosynthesis.

Despite all this, chemical modification even today has become more important and useful again for the determination of three-dimensional protein structures: It is applied with proteins too large for NMR spectroscopy and in cases where no crystals can be obtained for X-ray crystallography. In those cases the introduction of so-called reporter groups – fluorescence and spin labels – allows us to obtain information concerning the micro-environment of the group introduced by fluorescence or spin resonance (EPR spectroscopy). In addition one can estimate by these methods intramolecular distances and the shape and size of proteins.

Many biologically important proteins (receptors, transporters, ion channels, many enzymes) are membrane proteins, the crystallization of which is still difficult or (in many cases) impossible – despite significant progress achieved with, for example, potassium channels, calcium ATPases, rhodopsin and other G-protein coupled receptors, and several more. For these cases, three-dimensional information obtained by means of the introduction of reporter groups can be the method of choice.

A new development is the combination of chemical modification with site directed mutagenesis: by this one can introduce at appropriate positions of a polypeptide chain

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

functional groups for chemical modification (e.g., cysteine residues). These can be used to place reporter groups.

Another application of chemical modification is the investigation of quaternary structures and complexes with other proteins. Here the technique consists of the introduction of covalent links ("crosslinks") between subunits and neighboring molecules, respectively. These can be other proteins, nucleic acids, lipids, or peptides. Studies of this kind give clues as to spatial relationships. Application of crosslinking reagents of different length may allow estimates of distances between crosslinked molecules.

Finally, chemical modifications are essential tools for various practical applications: for example, for the synthesis of protein conjugates needed for antibody production, for designing material to be applied in affinity purification (affinity chromatography), for the detection of post-translational modifications of proteins, for radioactive labeling, and for the synthesis of fluorescently labeled proteins applicable in cell biology, especially in cytochemistry and histochemistry, in FRAP (fluorescence recovery after photobleaching) and FRET (Förster-resonance-energy-transfer).

FRET, Section 16.7

Chemical modifications can also be introduced into a protein by using chemically modified amino acids. This technology has been elaborated by Henry Lester. Proteomics and mass spectrometry use non-radioactive chemical modifications for quantitative analysis of proteins. For protein sequencing by mass spectrometry special end group modifications were introduced with the aim to reduce the complexity of fragmentation series. The most frequently used chemical modification was – until the introduction of Next Generation Sequencing – the modification with phenyl isothiocyanate, known as Edman sequencing.

Edman Sequencing, Section 14.1

Pharmaceutical applications of protein modification are, for example, modifications with polymers such as poly(ethylene glycol) to increase degradation during circulation, the introduction of functional groups for certain technological processes, and the introduction of crosslinks for the stabilization of proteins.

Proteome Analysis, Chapter 39
Mass Spectrometry, Chapter 15

Below we shall describe the principles of the applications of chemical modifications sketched above (as far as not described already in other chapters) with typical examples. Applications for the cleavage of polypeptide chains like the CNBr cleavage at methionine residues are described in Chapter 9. We shall mention only in passing reactions used for the removal of functional groups, which are tools for investigating the function of such groups. We shall focus on examples where chemical modifications are applied in structural analysis and practical application.

CNBr Cleavage, Section 9.6

6.1 Chemical Modification of Protein Functional Groups

In this chapter so-called side chain-specific reagents will be described. Most of the side chain reagents presently available are directed against the side chains of lysine, cysteine, tyrosine, histidine, methionine, arginine, and tryptophan residues, respectively. In some cases a reagent is specific for only one type of side chain. In many instances, in principle several of these chains can react. A typical example is the reaction of cysteine, histidine, methionine, and lysine residues with iodoacetamide. In this case the preferential reaction with only one type of side chain has to be secured by careful control of the reaction conditions, such as, for example, the pH value. Since functional groups of a given type can be located within a protein in different micro-environments, which affect the pK value of the dissociating group, specific groups can well react specifically with a reagent that cannot access the same group in their particular environment.

Lysine Residues The N-terminal α-amino groups and ε-amino groups of lysine residues are preferred targets of chemical modification. Lysine residues are ubiquitously present in proteins. They are located predominantly at the surface of a protein and therefore usually are accessible for modifying reagents. Reactions are performed at pH >9, when preferentially ε-amino groups are targeted. α-Amino groups have lower pK values than ε-amino groups and therefore can be modified selectively by lowering the pH. Amino group reagents are abundant.

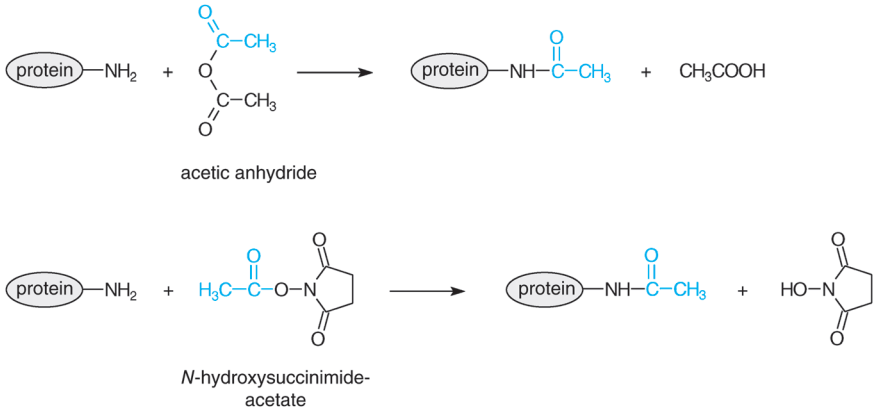


Figure 6.1 Acetylation of proteins with acetic anhydride or *N*-hydroxysuccinimide-acetate.

Acylation Acylation reactions are performed with anhydrides (e.g., acetic anhydride) or with active esters like *p*-nitrophenyl or *N*-hydroxysuccinimidyl ester (Figure 6.1).

Widespread application of radioactive labeling of proteins has been achieved with the Bolton–Hunter reagent (*N*-succinimidyl-3-(4-hydroxyphenylpropionate) (Figure 6.2). In a first step the reagent is radioactively iodinated, for example, with the Bolton–Hunter or the chloramine-T method. The mono- or di-iodinated reagent is used subsequently for alkylation of the protein (pathway 1 in Figure 6.2). Alternatively, the protein peptide is first substituted with the non-iodinated reagent. The derivative obtained can be stored and iodinated immediately before use (pathway 2). Of course, this method can be used only if no tyrosine or histidine residues are present as they would be iodinated as well.

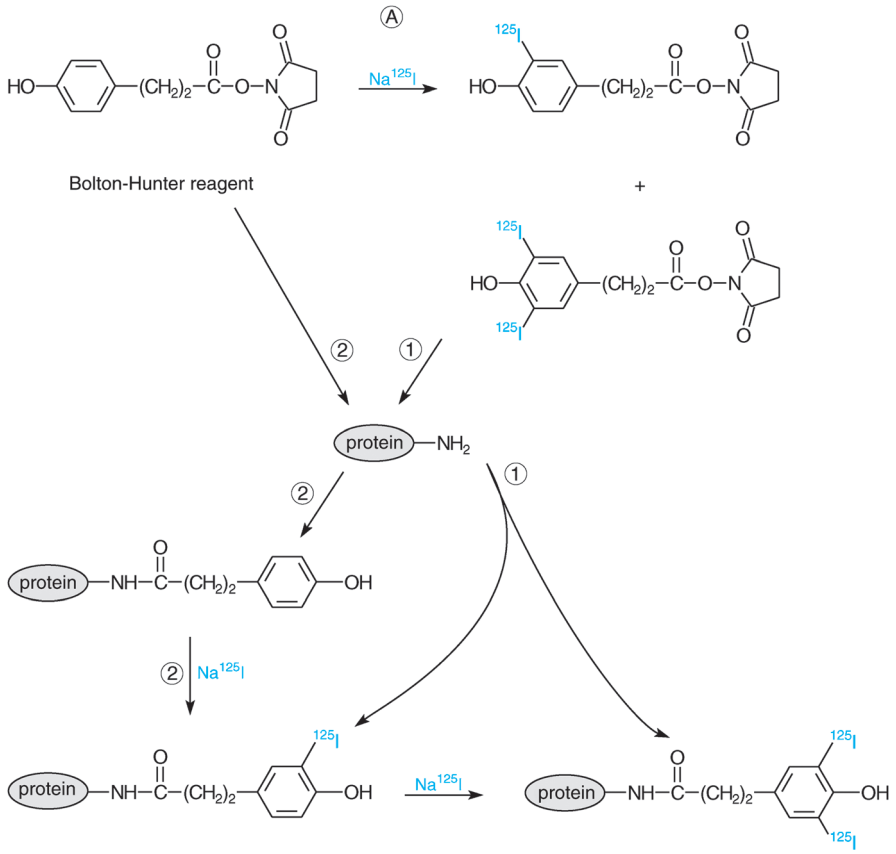


Figure 6.2 Radioactive labeling with the Bolton–Hunter reagent (*N*-succinimidyl-3 [4-hydroxyphenyl]propionate). The reagent can be iodinated first with $Na^{125}I$ and subsequently reacted with the protein to yield the radiolabeled protein derivative (arrow 1). Alternatively, the protein is alkylated with the uniodinated reagent. This is radiolabeled immediately before use.

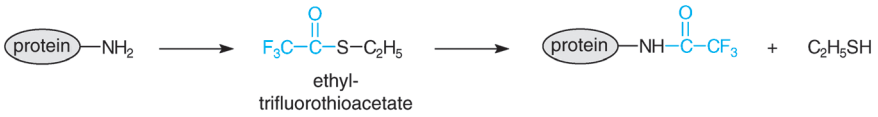


Figure 6.3 Introduction of a trifluoroacetyl group with ethyl trifluorothioacetate.

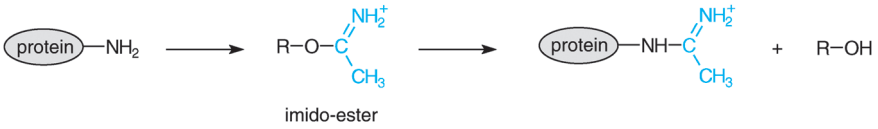


Figure 6.4 Amidation with an imido-ester.

Acylation removes positive charges and therefore alters the chromatographic properties of a protein considerably. By ion exchange chromatography or reversed phase HPLC mono-substituted products can be separated from unmodified or poly-substituted derivatives. Furthermore, the different mono-substituted derivatives can be separated because removal of positive charges at different sites having different micro-environments with different p*K* values renders proteins with different physicochemical and, thereby, different chromatographic properties.

Ion Exchange Chromatography, Section 10.4.7

The reaction with ethyl trifluorothioacetate (Figure 6.3) represents a mild method for the addition of a trifluoro-acetyl group. In this way various mono or multiple trifluoro-acetylated derivatives of small proteins like cytochrome *c* have been produced and investigated by ^{19}F NMR.

Amidination Modification of amino groups by means of imido esters (Figure 6.4) preserves the positive charge of lysine residues. A further advantage of this method is the solubility of imido esters in water, which react in the range pH 7–10 exclusively with ε- and α-amino-groups. Side reactions yielding modifications of methionine, tyrosine, and histidine residues are instable and decay rapidly.

Amidation is a convenient method for labeling proteins with ^{3}H or ^{14}C isotopes.

The modification can be removed if necessary by means of strong nucleophiles (hydrazine, hydroxylamine, aqueous methylamine).

Reductive Alkylation Amino groups react with aldehydes to yield Schiff bases that can be reduced with borohydrides to stable alkylamines (Figure 6.5). Most aldehydes form mono-substituted derivatives, reaction with formaldehyde results mainly in a dimethyl product.

Using for the reduction the, by far, mildest reagent $NaCNBH_3$ has the advantage that at neutral pH only the Schiff base and not, as with $NaBH_4$, the aldehyde itself is reduced. With NaB^3H_4 radioactive protein derivatives can be obtained. An important advantage of reductive alkylation is the preservation of the positive charge of the amino group. Methylation causes a slight increase of the p*K* value by about 0.5 units. In general, reductive alkylation does not affect the biological activity of the protein. It may lead to inactivation only if an essential lysine residue is involved.

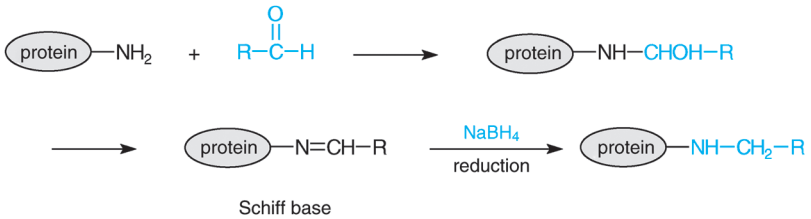


Figure 6.5 Formation of a Schiff base with aldehydes and subsequent reduction with sodium borohydride to yield a stable aminoalkyl compound.

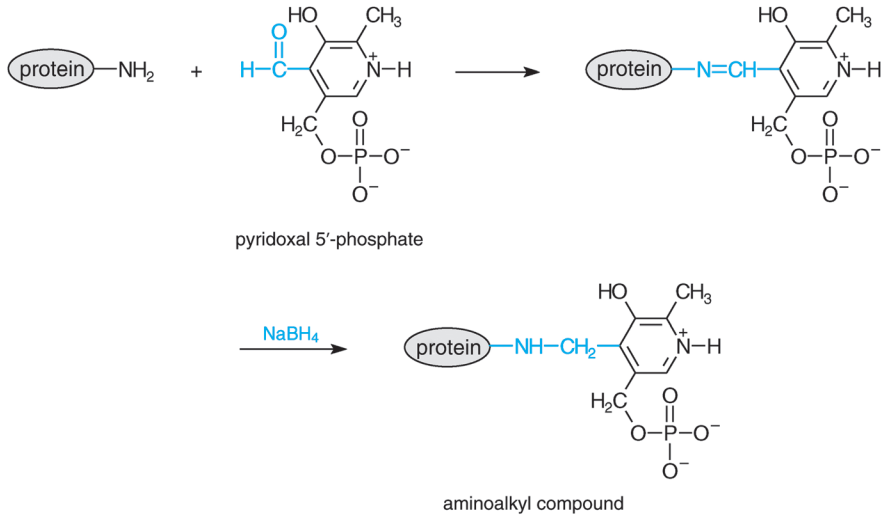


Figure 6.6 Formation of an aldimine with pyridoxal 5′-phosphate and subsequent reduction to yield an aminoalkyl compound.

Pyridoxal-5′-phosphate (Figure 6.6), the coenzyme of various amino transferases and decarboxylases, bound to an amino group of its respective apoenzyme as an aldimine is a fairly reactive aldehyde used for probing the accessibility of amino groups of water-soluble and membrane bound proteins, respectively.

Examples of naturally occurring (post-translational and non-enzymatic) aldehyde modified proteins are the retinal containing membrane bound chromoproteins: Special retinal isomers form Schiff bases with the ε-amino groups of lys216 and lys296 of bacteriorhodopsin and bovine rhodopsin, respectively. Reduction of these aldimines with $NaBH_4$ or $NaCNBH_3$ yields stable fluorescent products. In this way the binding locations in the proteins mentioned were determined.

Reaction with Isothiocyanate Phenyl isothiocyanate is used in classical peptide and protein sequencing. Various fluorescent groups can be introduced into proteins using their respective isothiocyanate derivatives.

Erdman Sequencing, Section 14.1

Caged Compounds A special development is the use of conjugates formed by proteins and so-called caged compounds. Caged compounds are temporarily inactive compounds that can be activated *in situ* at a given time by irradiation with light.

Caged Compounds, Section 7.1.5

The reagent 4,5-dimethoxy-2-nitrobenzyl chloroformate was introduced by Patchornik and coworkers for the solid phase synthesis of peptides as a light sensitive N-protecting group (Figure 6.7). Today the nitroveratryl oxycarbonyl group is used as a protecting group of α-amino groups in the solid state synthesis of peptide libraries on microchips. After the coupling step the protecting group is removed by irradiation.

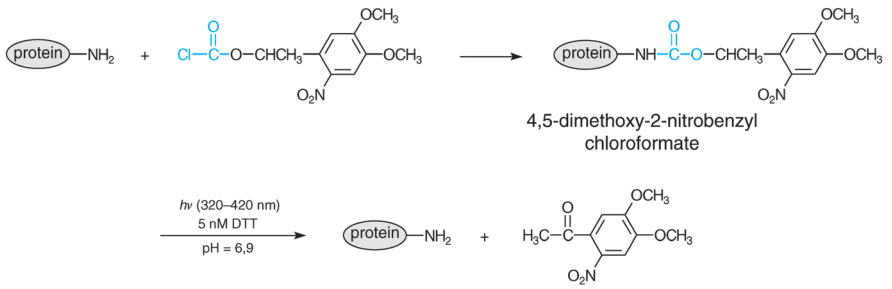


Figure 6.7 4,5-Dimethoxy-2-nitrobenzyl chloroformate, a modification protecting α-amino groups. The protecting group can be removed by irradiation.

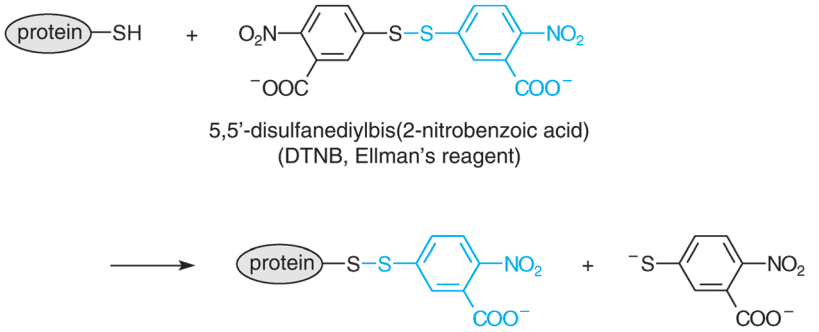


Figure 6.8 Modification of SH-groups with Ellman's reagent.

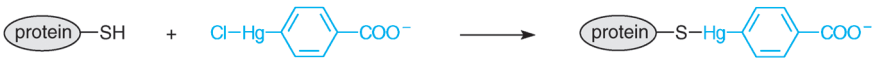


Figure 6.9 Analysis of SH-groups with *p*-chloromercury benzoate.

Cysteine Residues A large repertoire of modification reactions is available for protein SH groups. For their detection and quantification the Ellman reagent is still the method of choice (Figure 6.8). The thionitrobenzoate released can, because of its large absorption coefficient (ε_{412} 13 600 at pH 8.00), easily be determined spectroscopically.

In classical enzymology –SH groups were determined by means of *p*-chloromercury benzoate, pCMB (Figure 6.9).

SH groups react with iodo-acetate and iodo-acetamide and also with maleimides (Figure 6.10). Many methods for introducing groups for spectroscopy are based on these two reactions.

Light sensitive groups can be introduced into proteins not only via their amino groups but also via SH-groups of cysteine residues. For this purpose 2-nitrobenzyl bromide or the water soluble 2-nitrohydroxybenzyl bromide is applied (Figure 6.11). The attachment of haptens to SH-groups can be useful for antibody production.

Chemical modification of other protein functional groups is touched upon here only briefly, because they are applied only rarely as compared to amino- and SH-group modifications.

Glutamate and Aspartate Residues Carboxy groups of side chains of these residues as well as carboxy terminal groups can be modified with carbodiimides (Figure 6.12), especially with water soluble derivatives as, for example, 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide. HX represents here a nucleophile; when X is an amino group of the protein itself an intramolecular crosslink results (or an intermolecular one, when carboxy- and amino group are contributed by different proteins). A practical application of such crosslinking reactions is the

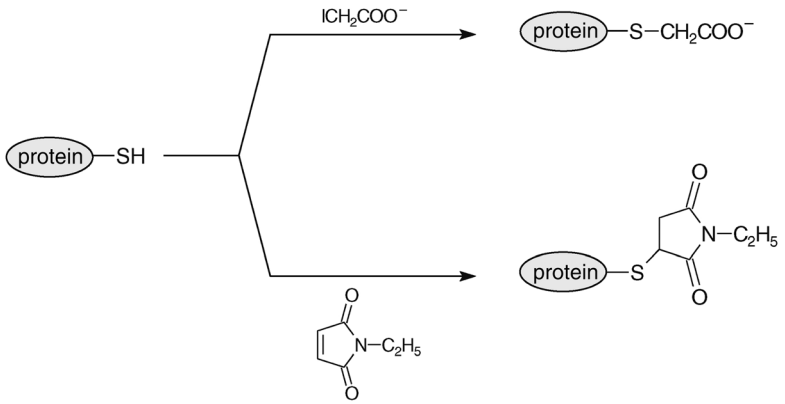


Figure 6.10 Reaction of proteins with iodoacetate or *N*-ethylmaleimide.

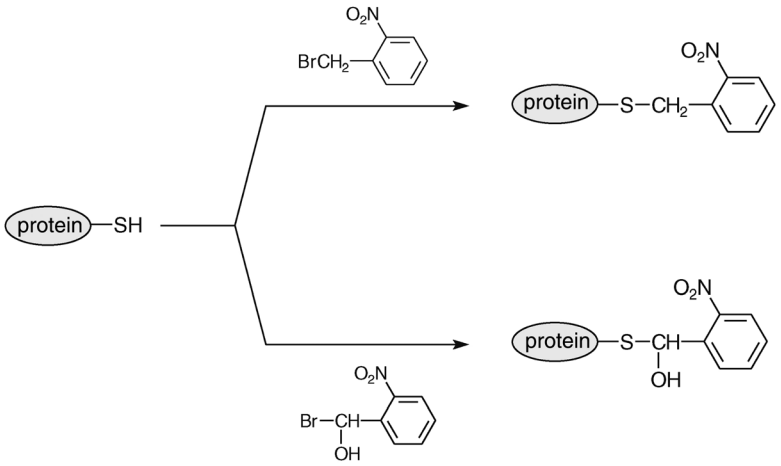


Figure 6.11 Modification of proteins with 2-nitrobenzyl bromide and 2-nitro-hydroxybenzyl bromide.

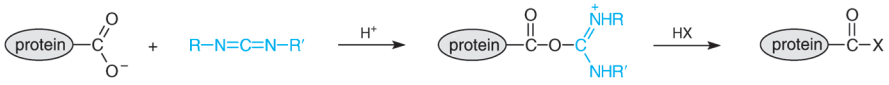


Figure 6.12 Modification of a C-terminal carboxy group with a carbodiimide.

production of protein–protein or protein–peptide conjugates, for example, for the production of antibodies against certain peptides. Adding an external HX nucleophile like N-substituted glycinamide yields radioactive or spectroscopically visible derivatives.

Arginine Residues The guanidinium groups of arginine residues can, for the purpose of determining their functional role in the active center of an enzyme or in the ligand binding pocket of a receptor, be fairly specifically modified with various 1,2- or 1,3-dicarbonyl compounds (glyoxal, phenylglyoxal, 2,3-butanedione, etc.). Figure 6.13 illustrates the reaction with phenylglyoxal.

Tyrosine Residues Radioactive labeling of proteins by iodination of tyrosine residues with chloro amine T is an indispensable tool in many biological fields (Figure 6.14). This reaction yields mono- as well as di-iodinated tyrosine derivatives. Iodination usually does not destroy a protein's activity (except in cases where an "essential" tyrosine residue is involved). Iodides like $Na^{125}I$, having specific radioactivities of up to 2000 Ci $mmol^{-1}$, are sources of radio-activity. For reaction with the protein the iodides are converted into highly reactive species (I_2 or ICl) by means of strong oxidants like chloroamine T, iodogen, or iodo-beads. For the same purpose one can also apply milder conditions: catalyzed by the enzyme lactoperoxidase in

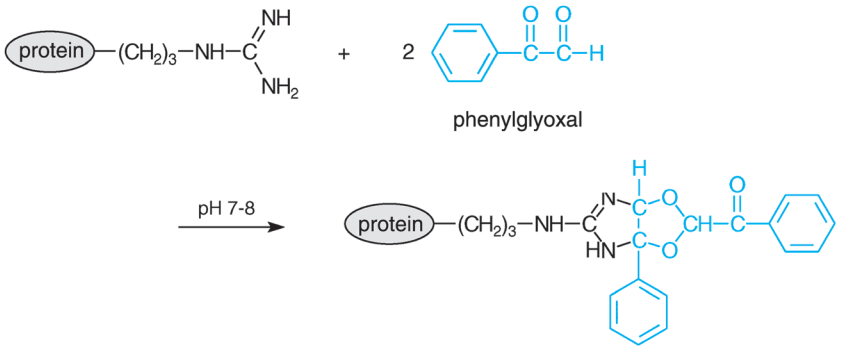
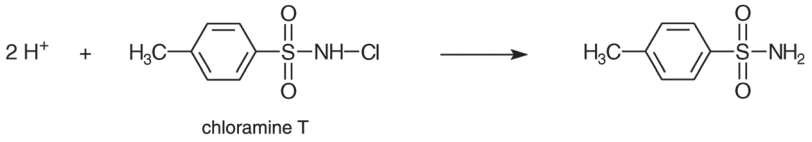


Figure 6.13 Reaction of arginine resi-dues with phenylglyoxal.



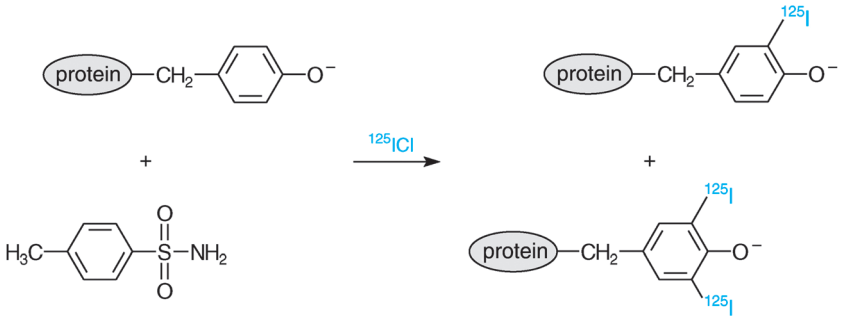


Figure 6.14 Iodination of tyrosine residues by chloramine T.

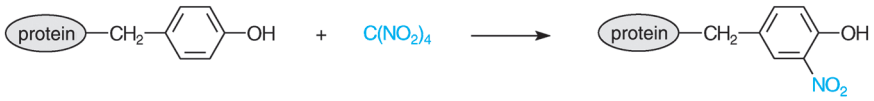


Figure 6.15 Modification of tyrosine residues with tetranitromethane.

the presence of H_2O_2, I^- can be converted into I_2. If a protein contains accessible histidine residues these may iodinated as well.

Frequently, tyrosine residues of proteins are modified by tetranitromethane (Figure 6.15). The 3-nitrotyrosine group seems to be a useful reporter group: at low pH it has an absorption maximum around 360 nm ($\varepsilon_{360} \approx 2800\,M^{-1}\,cm^{-1}$), which is shifted at higher pH values to 428 nm, accompanied by an increased absorption ($\varepsilon_{428} \approx 4200\,M^{-1}\,cm^{-1}$). 3-Nitrotyrosine residues possess a pK value of ca. 7.0. Therefore, this modification affects the ionization of a tyrosine, the functional role of which can be investigated this way. Depending on the environment within the protein the pK value is about 10.5. Because the pK value of 3-nitrotyrosine is sensitive to its micro-environment, this group can deliver valuable information as to the milieu in the neighborhood of the group modified. A serious limitation though of the tyrosine modification is that cysteine, methionine, and tryptophan residues can be modified likewise by the reagent. Furthermore, it can cause intra- and intermolecular crosslinks.

Interestingly, nitration of tyrosine residues can be caused by endogenously occurring NO, thereby possibly being involved in certain neurodegenerative diseases. NO also can lead to S-nitrosylation of cysteine residues.

Tryptophan Residues Modification of tryptophan residues usually aims to cleave the protein at the C-terminal side of the tryptophan. Reagents used for this purpose are *o*-iodosobenzoic acid or 2-(2-nitrophenylsulfonyl)-3-bromo-3-methylindolenine (BNPS-skatole). Sometimes bromosuccinimide is used to probe the effect of tryptophan oxidation on the activity of the protein. *N*-Bromosuccinimide can oxidize tyrosine and histidine residues as well and can cleave the peptide bond on their C-terminal side. This limits the interpretation of investigations of the function of tryptophan residues using *N*-bromosuccinimide oxidation.

Oxidative Cleavage at Tryptophan Residues, Figure 9.6.

Tryptophan residues can also be modified with 2-hydroxy-5-nitrobenzyl bromide (Koshland's reagent (Figure 6.16). Sulfenyl halogenides, which react with SH groups, can modify tryptophan residues, too (Figure 6.17).

Methionine Residues The most common and practicable application of methionine modification is the cleavage of proteins C-terminal to methionine with CNBr. Methionine residues can also be the predominant site of modification with iodoacetamide or iodoacetic acid,

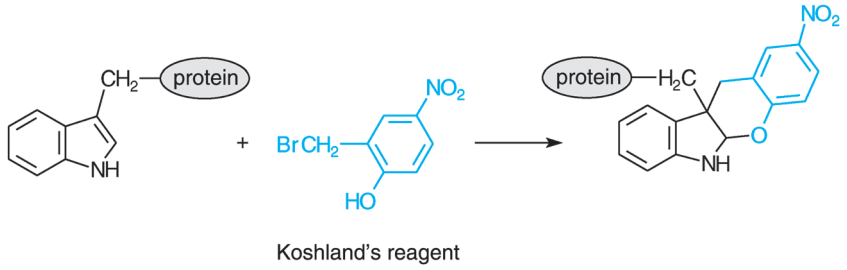


Figure 6.16 Modification of tryptophan residues with 2-hydroxy-5-nitrobenzyl bromide (Koshland's reagent).

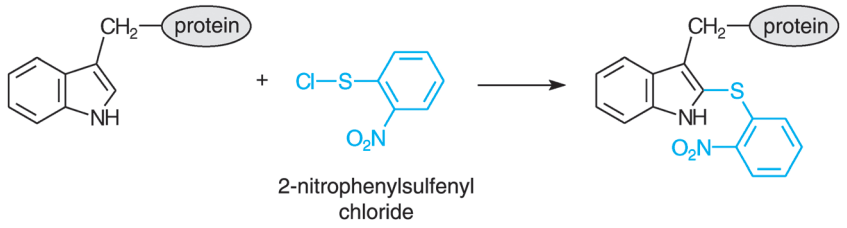


Figure 6.17 Modification of tryptophan residues with 2-nitrophenylsulfenyl chloride.

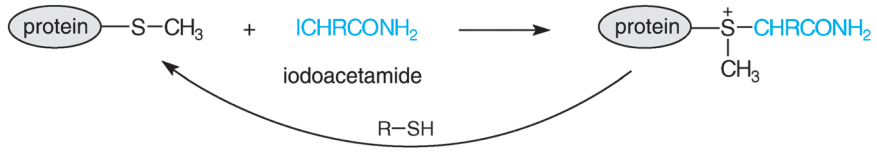


Figure 6.18 Modification of methionine residues with derivatives of iodoacetamide.

if the reaction is performed at pH 4–5 or below (Figure 6.18). In this way, various reporter groups can be introduced. An advantage of this reaction is its reversibility.

CNBr Cleavage, Section 9.6

Histidine Residues Diethyl pyrocarbonate (Figure 6.19) is a reagent with high selectivity for histidine residues. It is used for probing the functional role of histidine residues in a protein.

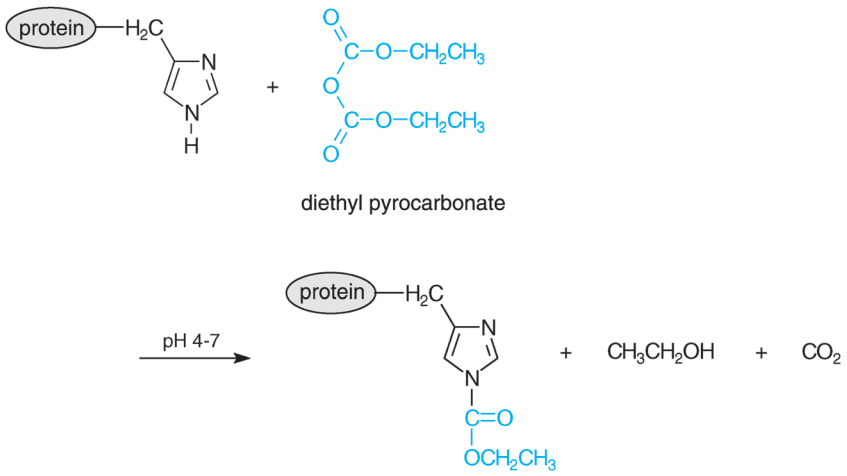


Figure 6.19 Modification of histidine residues with diethyl pyrocarbonate (diethyl dicarbonate).

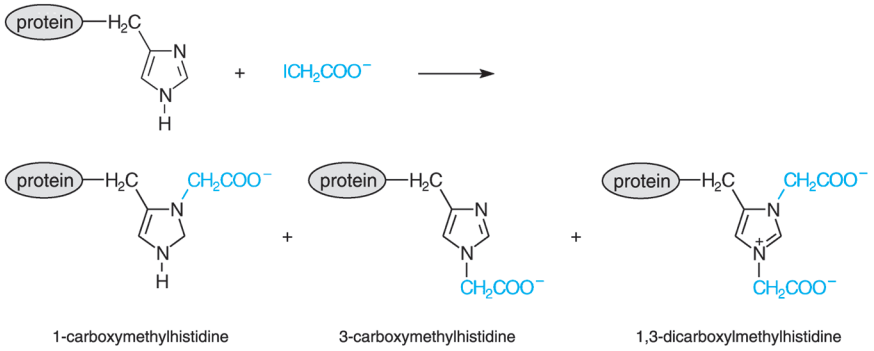


Figure 6.20 Modification of histidine residues with iodoacetate to yield 1- and 3-carboxymethylhistidine and 1,3-dicarboxylmethylhistidine.

This is important since many enzymes like RNAses and serine proteases possess histidine residues in their active center.

Iodoacetate and iodoacetamide, which we discussed already as reagents for SH-, amino-, and methionine groups, modify histidine residues, too. The reaction can result in 1- and 3-carboxymethylhistidine and in disubstituted 1,3-dicarboxymethylhistidine (Figure 6.20).

A special case is "affinity modification:" The modifying reagent contains a reactive group and a structure with high affinity for the active site of an enzyme or for the ligand-binding pocket of a receptor. The reactive group interacts with a functional group of the enzyme or receptor, respectively, in the vicinity of the molecular structure supposed to mediate the docking of the reagent to the binding site. The affinity group targets the reactive group to its site. In this way selectivity is achieved: The active site accumulates the reagent. Consequently, it reacts with the active site much more rapidly than elsewhere. Classical examples are the halogen ketones *N*-tosyl-L-phenylalanine-chloromethyl ketone (TPCK) and *N*-tosyl-L-arginine-chloromethyl ketone, which modify selectively histidine residues in the active site of chymotrypsin and trypsin, respectively.

6.2 Modification as a Means to Introduce Reporter Groups

6.2.1 Investigation with Naturally Occurring Proteins

Fluorescence Labeling Incorporation of fluorescent groups in proteins by chemical modification serves different purposes and therefore must be adapted to different requirements. For the determination of 3D structures of proteins it is necessary to know the number and location of the groups introduced. For practical purposes, on the other hand, such as, for example, the production of fluorescent antibodies, the modified protein just has to fluoresce – the more the better.

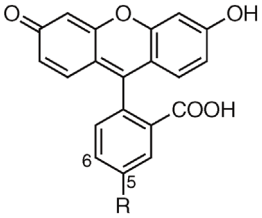
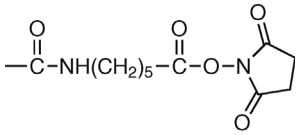
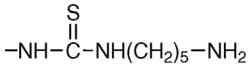
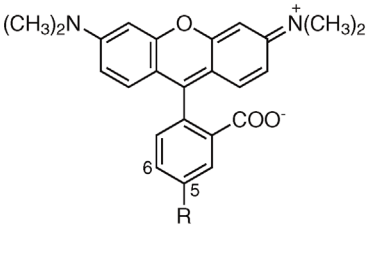
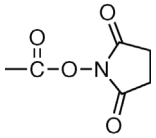
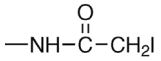
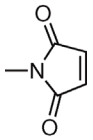
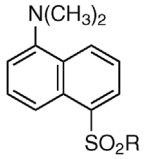
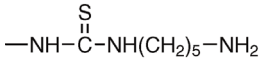
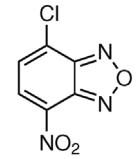
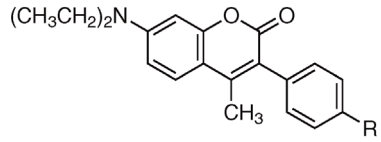
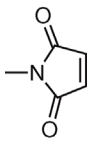
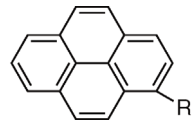
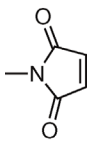
Fluorescence Spectroscopy, Section 7.3

The introduction of fluorescent groups uses the same reactive groups described above (Section 6.1). Preferred reaction partners are amino- and SH-groups. Succinimidyl ester or isothiocyanates are used for the attachment to amino groups (Table 6.1). Most fluorescence labelings result in derivatives of fluorescein and rhodamine, respectively. Their spectral characteristics are listed in Table 6.1. Fluorescein isothiocyanate (FITC, Figure 6.21) and tetramethylrhodamine isothiocyanate are most commonly used for the production of fluorescent antibodies.

Sulfonyl chloride and dansyl chloride (Figure 6.22) are used for fluorescent labeling, too. As well as with amino groups they react with tyrosine, histidine, and cysteine residues. Compared to the sulfonamides the respective derivatives are less stable.

Dansyl amides are also stable under the conditions of amino acid analysis. Dansyl chloride, therefore, can be used for the determination of the amino terminal amino acid of peptides and proteins and for the dansyl variant of Edman degradation.

Table 6.1 Reagents for introducing fluorophores.

1 2 3			fluorescein-5-isothiocyanate
			6-[fluorescein-5(6)-carboxamido]hexanoic acid *N*-hydroxysuccinimide ester
			fluorescein cadaverine
4 5 6 7			tetramethylrhodamine-5-isothiocyanate
			5-carboxy-tetramethylrhodamine N-succinimidyl ester
			tetramethylrhodamine-5-(6)-iodoacetamide
			tetramethylrhodamine-5-(6)-maleimide
8 9			dansyl chloride (DNSCl)
			dansylcadaverine
10		—	4-chloro-7-nitrobenz-2-oxa-1,3-diazole (NBD-chloride)
11			7-diethylamino-3-(4-maleimidophenyl)-4-methylcoumarin
12			*N*-(1-pyrenyl)maleimide

The optical properties of dansyl derivatives depend on their micro-environment: a shift of the emission maximum to shorter wavelengths and an increase in fluorescence indicate a hydrophobic environment. Dansyl residues are favorable partners in acceptor–donor pairs with tryptophan residues as intrinsic fluorophores. Therefore, they are useful for the determination of distances by means of fluorescence energy transfer.

Fluorescence Spectroscopy, Section 7.3

To this end a dansylated ligand is bound to its binding site, for example, to the active site of an enzyme. Irradiation with light of the appropriate wavelength results in a transfer of part of its

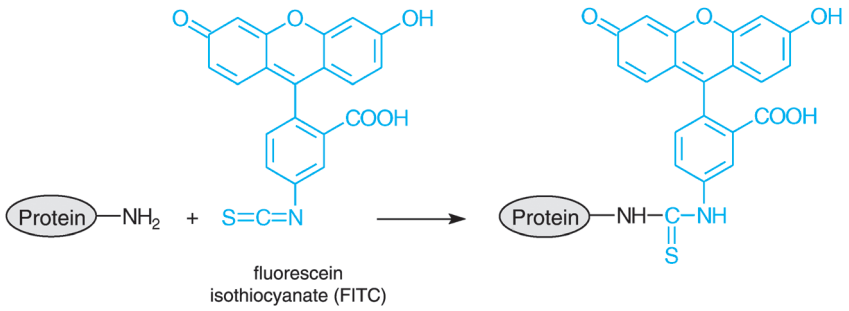


Figure 6.21 Fluorescence labeling with fluorescein isothiocyanate (FITC).

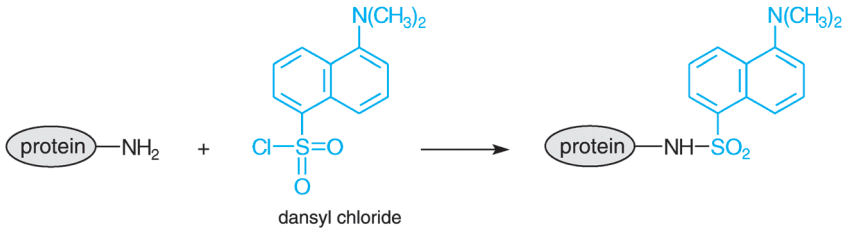


Figure 6.22 Fluorescence labeling with dansyl chloride.

energy to a neighboring tryptophan residue, which in turn is excited to fluoresce. The yield of this fluorescence, relative to the energy irradiated, correlates with the distance (and orientation) of the dansyl and tryptophan residues. Another reagent applied for spectroscopic studies with proteins is 4-chloro-7-nitrobenz-2-oxa-1,3-diazole (NBD chloride) (Table 6.1, formula 10). It reacts with NH_2- and SH-groups. The resulting fluorescence depends largely on the nature of the group modified and on the polarity of the solvent.

Besides dansylation, two other reactions that link fluorescent groups with amino groups of proteins permit the sensitive detection of proteins and must not be omitted in this context: reaction with fluorescamine and with *o*-phthaldialdehyde (OPA), respectively. Fluorescamine itself does not fluoresce, but it forms with amino groups fluorescent reaction products (Figure 6.23). These products are useful for the detection and quantification of proteins. o-Phthaldialdehyde (OPA) reacts with amino groups in the presence of thiols, forming intensively fluorescent isoindoles (Figure 6.24). This reaction allows the detection of proteins in the picomole range and therefore is used for amino acid analysis.

Modification with OPA, Section 13.3

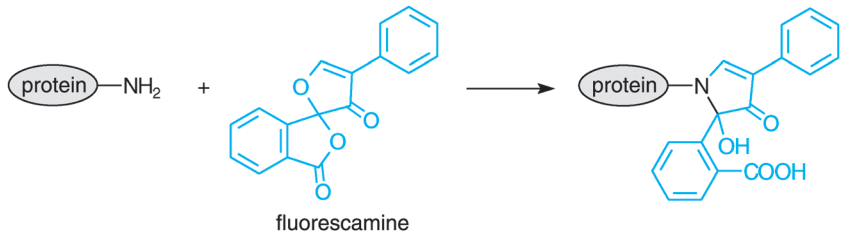


Figure 6.23 Fluorescence labeling with fluorescamine.

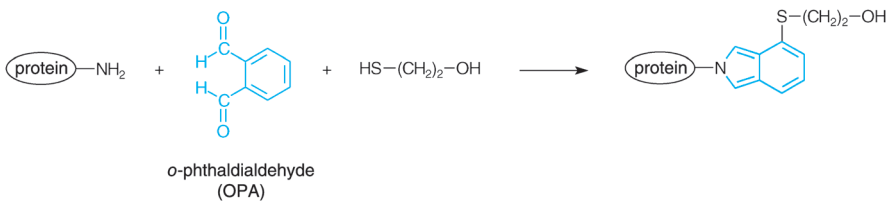


Figure 6.24 Fluorescence labeling with phthaldialdehyde in the presence of thiols.

Fluorescent groups can be introduced into cysteine residues of proteins via their iodoacetamide or maleimide derivatives. As shown in Table 6.1, again fluorescein or rhodamine groups are the preferred fluorophores. Coumarins (no. 11 in Table 6.1) are also used for fluorescent labeling of proteins. Coumarins and fluorescein are also good donor–acceptor pairs for energy transfer measurements.

The maleimides used for the labeling reactions usually are not fluorescent by themselves. Fluorescence results only upon covalent attachment to an -SH group of a protein. A reagent of this kind is pyrene maleimide (12, Table 6.1). It reacts preferentially with residues in hydrophobic pockets. Analysis of fluorescence quenching by fatty acids of a lipid membrane, which contain a spin label (see below) in different positions, allows an estimate of the position of the fluorophore relative to the membrane–water interface.

The reason why fluorescein cadaverine and dansyl cadaverine (3 and 9, respectively, Table 6.1) are considered here is that sometimes fluorescent groups can be introduced into a protein not by chemical modification but rather by a transglutaminase catalyzed reaction. For this dansyl- and fluorescein-cadaverine are useful substrates (Figure 6.25).

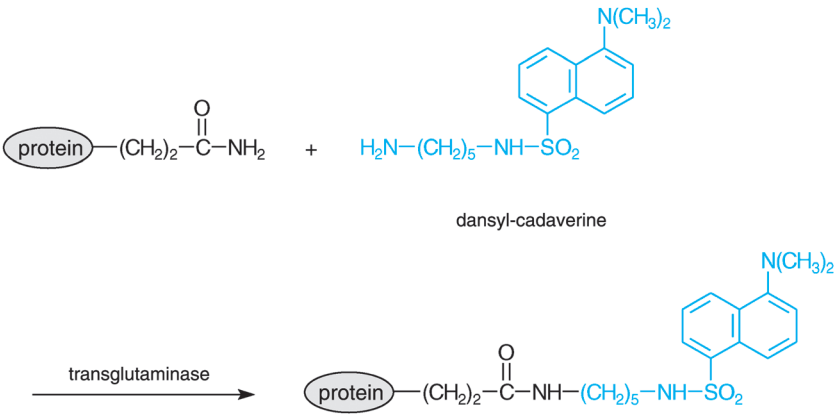


Figure 6.25 Fluorescence labeling by a reaction with dansyl-cadaverine, catalyzed by a transglutaminase.

Spin Label Stable compounds with an unpaired electron, for example, the N-oxides of tetramethylpyrrolidine- or tetramethylpiperidine derivatives (Table 6.2) allow measurements of electron paramagnetic resonance (EPR) spectra. Covalently or non-covalently bound they are called "spin labels".

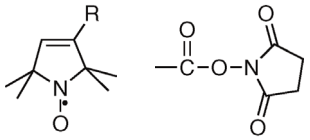
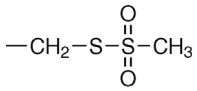
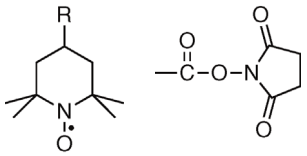
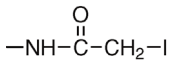
From EPR spectra the mobility of a spin label at is site of attachment can be deduced. The spin label reports on the micro-environment at its location. One application is to determine whether the label is located at the surface or in the interior of a protein. To this end the accessibility by a paramagnetic ion (e.g., Ni^{2+}) added externally is measured. If they collide this is indicated by a certain broadening of the signal in the EPR spectrum. Furthermore, spin labels can also be used together with fluorescent labels, because they quench their fluorescence, when they come close together. Methods for the introduction of spin labels are similar to those described for fluorescence labels. Here, too, modification of amino and –SH groups is easiest. Some typical succinimide and iodoacetamide spin label derivatives are listed in Table 6.2. If no –SH groups are present labeling of histidine residues with iodoacetamides can be achieved.

EPR, Section 18.2

Originally, the problem with spin labeling of proteins was that only rarely could a label be introduced at a specified position. A successful example for this was a series of mono-labeled α-neurotoxins from snake venoms and for EGF (epidermal growth factor) labeled in only one position (a lysine residue). These derivatives were useful tools in the investigation of ligand–receptor interactions.

Two spin labels located at specified positions of a protein can be used for estimating distances between such labels by measuring their paramagnetic dipole interaction. Because of technical problems this method was rarely used for naturally occurring proteins. But the method has become useful with recombinant and mutated proteins (see below).

Table 6.2 Reagents for spin labeling.

1		succinimidyl-2,2,5,5-tetramethyl-3-pyrroline-1-oxyl-3-carboxylate
2		*S*-(1-oxyl-2,2,5,5-tetramethyl--pyrroline-3-methyl) methanethiosulfonate
3		succinimidyl-2,2,6,6-tetramethylpiperidine-1-oxyl-4-carboxylate
4		2,2,6,6-tetramethyl-4-(2-iodacetamid)-piperidine-1-oxo

6.2.2 Investigation of Recombinant and Mutated Proteins

Recombinant DNA technologies have changed classical protein chemistry significantly. Heterologously overexpressed and at specific sites mutated proteins can be produced rapidly, avoiding lengthy purification and chemical modification procedures. Especially, site directed mutagenesis by genetic engineering makes proteins available that contain only one site available for a special modification. This concept is the basis for a technique called the scanning cysteine accessibility method (SCAM), which was developed for the investigation of structure–function relationships in membrane-bound receptors and other proteins.

For example, photoaffinity labeling and mutagenesis/chimera experiments led to the assumption that the ion channel of the nicotinic acetylcholine receptor is formed by the five transmembrane helices M2, each one contributed by one of the five receptor subunits. To specify the side chains participating directly in the ion conductance, Karlin and coworkers substituted one by one the M2 amino acids by cysteine. These modifications influence the channel properties only slightly. By reacting the respective cysteine residue with a charged hydrophilic reagent the channel was blocked. This block was achieved only with cysteine residues located in the center part of the M2 sequence and only with every other position. This periodicity indicated that a ß-strand is involved. An α-helical structure would have caused a block at only every fourth position.

A similar series of cysteine residues introduced by site-directed mutagenesis can be used, too, for the introduction of spin labels for EPR analysis. In this manner, time dependent conformational changes, the micro-environment of the label, and intra- and intermolecular distances can be determined. At present, site specific spin labeling is based largely on cysteine-scanning mutagenesis. The method is recommended for proteins that are not yet crystallizable or accessible to NMR spectroscopy. An example is bacteriorhodopsin, where each position of the trans-membrane helices D and F was, one by one, substituted by a cysteine and subsequently modified with a nitroxy group (Table 6.2) (nitroxy scanning). In this way, folding models deduced from electron microscopy could be verified. In some important cases such results were confirmed by X-ray crystallography, proving the value of these methods.

Intramolecular distances between covalently bound nitroxides and specified groups of protein metal ion complexes can also be measured. This, too, can be achieved by site-directed mutagenesis, introducing the metal binding group at a specific position.

6.3 Protein Crosslinking for the Analysis of Protein Interaction

Crosslinking reagents are used to determine spatial relationships within a protein, but more frequently to investigate interactions between proteins or proteins and nucleic acids or lipids. In addition to the criteria mentioned above for the selection of an appropriate reagent for chemical modification, in crosslinking several additional points have to be observed: Besides the specificity and selectivity it is important to know whether a reagent can permeate a membrane and whether the distance between the groups to be crosslinked is favorable. Furthermore, in many cases the use of cleavable crosslinkers facilitates the analysis of the crosslinking products.

Two applications of crosslinking shall be emphasized: Measurement of distances by means of crosslinkers of different length, and the analysis of "neighborhoods", the detection of molecules in close proximity of a protein that might be functional partners.

This chapter describes crosslinking with bifunctional reagents. Below we shall turn to photo-crosslinking. Basically, every photoaffinity labeling is a crosslinking reaction, though only some specific applications are considered in this context. We shall see that differentiation between photoaffinity labeling and photo-crosslinking is not always possible.

6.3.1 Bifunctional Reagents

Bifunctional crosslinking reagents come in three groups: homo-bifunctional, hetero-bifunctional, and zero length reagents. Examples of the latter are the formation of an amide bond between amino- and carboxyl groups of a protein using a carbodiimide, and the formation of disulfide bonds between cysteine -SH groups. Homo-bifunctional reagents (Table 6.3) carry two functional groups separated by a spacer of varying length. Examples are 1,5-difluoro-2,4-dinitrobenzene (no. 1 in Table 6.3) and formaldehyde and glutaraldehyde (2 and 3 Table 6.3), *N*-hydroxysuccinimido esters (4–8), and the imidates (9–12). DSS (no. 5) is often used for crosslinking radioactively labeled peptides or proteins with a protein. Activated esters (4 and 5) and imidoesters (9–11) are not cleavable in themselves, but cleavable derivatives are commercially available. Cleavage is achieved using reducing agents (6 and 12) or by oxidizing diols with neighboring –OH groups (7 and 8 in Table 6.3).

Homobifunctional reagents are convenient for the determination of molecular weights and quaternary structures of proteins composed of more than one polypeptide chain. Covalent crosslinking prevents protein complexes from falling apart, rendering the molecular weight of the complex. Occasionally the *crosslink* can be traced down to the reacting amino acids. This allows important information to be obtained concerning the structure and neighboring molecular entities.

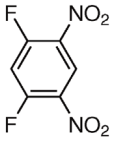
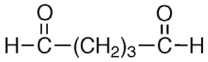
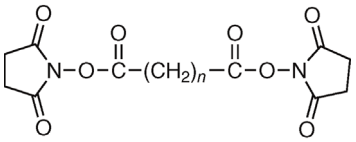
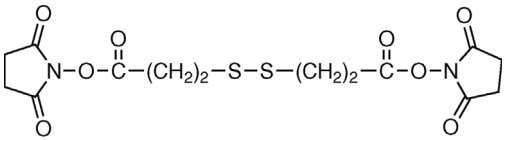
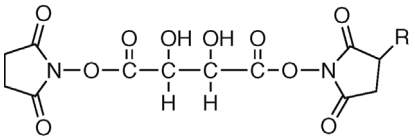
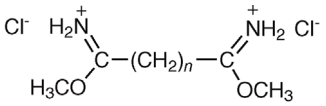
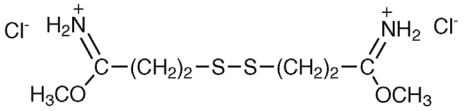
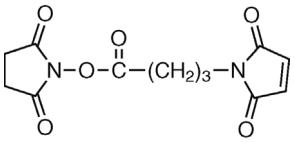
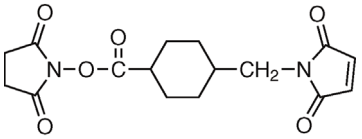
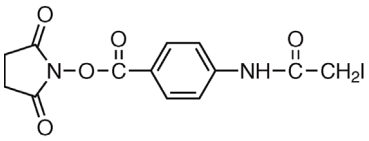
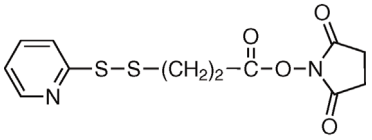
Hetero-bifunctional reagents possess two different reactive groups, for example, a maleimido or iodoacetamido group, for the reaction with a –SH group, on the one hand, and an activated ester for the reaction with an amino group of a protein on the other (13–15 in Table 6.3). Reagent number 16 in the table is a further example of a cleavable bifunctional reagent. But here the reductive cleavage serves a special purpose. It is used not as usual for the reductive cleavage but for the removal of a protection group and by that for the release of a –SH group for the connection with another –SH group, or for its modification with a specific –SH reagent.

6.3.2 Photoaffinity Labeling

This method uses a ligand, for example, a substrate of an enzyme or an agonist or antagonist of a receptor, derivatized with a group that can be activated by light. Similar to the situation with chemical modification the affinity of the ligand brings the reactive group to its target. The main difference is that the group is activated only after it arrives at its target. The activation by light allows us to determine the site and time of the labeling reaction. This makes photoaffinity reagents more selective than other covalent labels, at least in theory. Applications extend from simple labeling of a protein in a protein mixture to detailed mapping of binding sites.

Photochemistry today applies very different types of photo-reactions: In the late 1970s and early 1980s aryl azides were favored that form nitrenes upon photo-activation. Their main advantage is that they can be prepared most easily. It was assumed that nitrenes insert into many different bonds including C—H bonds and therefore may "hit" any amino acid residues within their neighborhood. Aryl azides served their purpose as long as they were aimed at identifying a macromolecular target, an enzyme, receptor, transporter, any binding protein. As one attempted to identify the exact reaction site within its target significant problems occurred: First of all the yield of the photoreactions was low, in most cases not higher than a few percent. Furthermore,

Table 6.3 Bifunctional reagents.

1		1,5-difluoro-2,4-dinitrobenzene
2		formaldehyde
3		glutaraldehyde
4 5		disuccinimidyl glutarate disuccinimidyl suberate
6		dithiobis(succinimidyl propionate)
7 8		disuccinimidyl tartrate disulfosuccinimidyl tartarate
9 10 11		dimethyl adipimidate · 2 HCl dimethyl pimelimidate · 2 HCl dimethyl suberimidate · 2 HCl
12		dimethyl 3,3′-dithiopropionimidate · 2 HCl
13		*N*-γ-maleimidobutyryl-oxysuccinimide ester
14		succinimidyl 4-(N-maleimidomethyl)cyclohexane-1-carboxylate
15		*N*-succinimidyl (4-iodoacetyl)aminobenzoate
16		succinimidyl 3-(2-pyridyldithio)propionate

the lifetime of the activated entity was long. Consequently, non-specific or multiple labelings took place. Therefore, diazirines seemed more promising (11–17, Table 6.4). Upon photolysis they form carbenes. Their half-life is much shorter, which reduces their "promiscuity". As with the nitrenes the yield is low, too. This situation is exacerbated by the fact that the products formed are unstable in many cases and do not survive isolation and characterization by methods of classical protein chemistry.

Much higher crosslinking yields (70–80%) are achieved using benzophenone derivatives, for example, *p*-benzoyl-L-phenylalanine (Bpa, no. 20 in Table 6.4).

Table 6.4 Reagents for photolabeling.

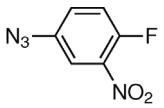
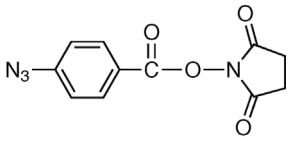
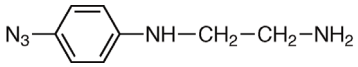
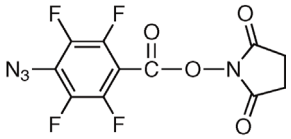
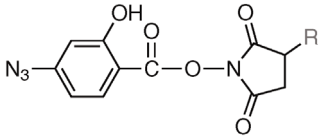
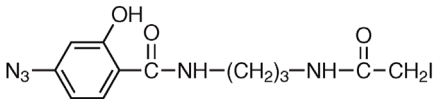
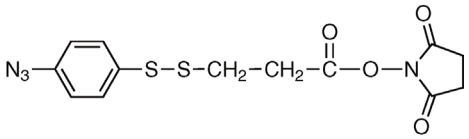
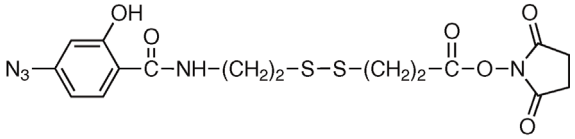
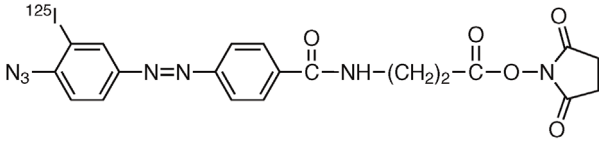
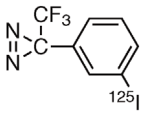
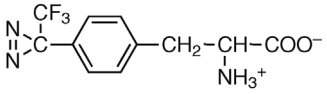
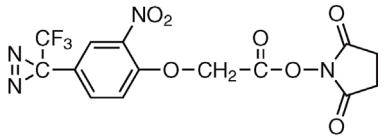
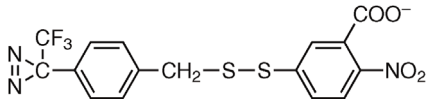
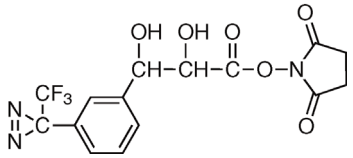
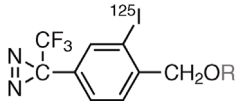
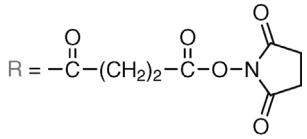
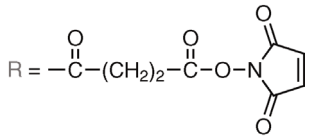
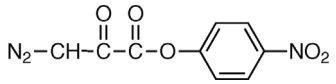
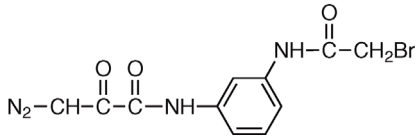
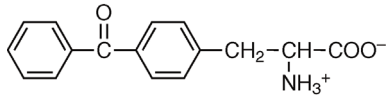
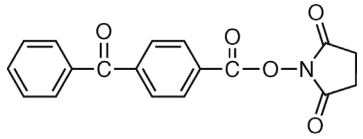
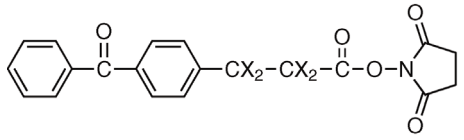
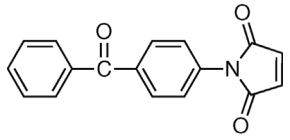
1		4-fluoro-3-nitrophenyl azide
2		4-azidobenzoic acid N-hydroxysuccinimide ester
3		*N*-(β-aminoethyl)-4-azido-2-nitroaniline
4		*N*-succinimidyl 4-Azido-2,3,5,6-tetrafluorobenzoate
5 6	R = H R = SO_3Na 	*N*-hydroxysuccinimidyl-4-azidosalicylic acid *N*-hydroxysulfosuccinimidyl-4-azidosalicylic acid
7		1-(p-azidosalicylamido)-4-(iodoacetamido) butane
8		*N*-hydroysuccinimidyl-4(4-azidophenyl) 1,3′dithiopropionate
9		sulfosuccinimidyl-2-(p-azidosalcylamido)ethyl-1,3-dithiopropionate
10		*N*-((3-((4-((4-azido-3-iodophenyl)azo)benzoyl)amino)propanoyl)-oxy)succinimide (Denny-Jaffe reagent)
11		3-(trifluoromethyl)-3-(m-[^{125}I]iodophenyl)diazirine ([^{125}I]TID)
12		4-(trifluoromethyl-diazirine)phenylalanine

Table 6.4 Reagents for photolabeling (Continued).

13		[2-nitro-4-(3-(trifluoromethyl)-3H-diazirin-3-yl)phenyl]acetyl-*N*-hydroxy-succinimide ester
14		
15		3-[3-(3-(trifluoromethyl)diazirine-3-yl)phenyl]-2,3-dihydroxypropanoic-*N*-hydroxysuccinimide ester
		
16		3-[[[2-([^{125}I])iod-4-[3-(trifluoromethyl)-3H-diazirin-3-yl]benzyl]-oxy]-carbonyl]propanoyl-*N*-hydroxysuccinimide ester
17		2-([^{125}I])Iod-4-[3-trifluoromethyl)-3H-diazirine-3-yl]benzyl-3-maleimidopropionic ester
18		4-nitrophenyl 3-diazopyruvate
19		*N*-bromoacetyl-*N*′-(3-diazopyruvoyl)-rn-phenylalanine
20		*p*-benzoyl-L-phenylalanine
21		*p*-benzoyl benzoyl-*N*-hydroxysuccinimide ester
22		[^{3}H]p-benzoyl dihydroxycinnamoyl-*N*-succinimide ester
23		4-(maleimido)benzophenone

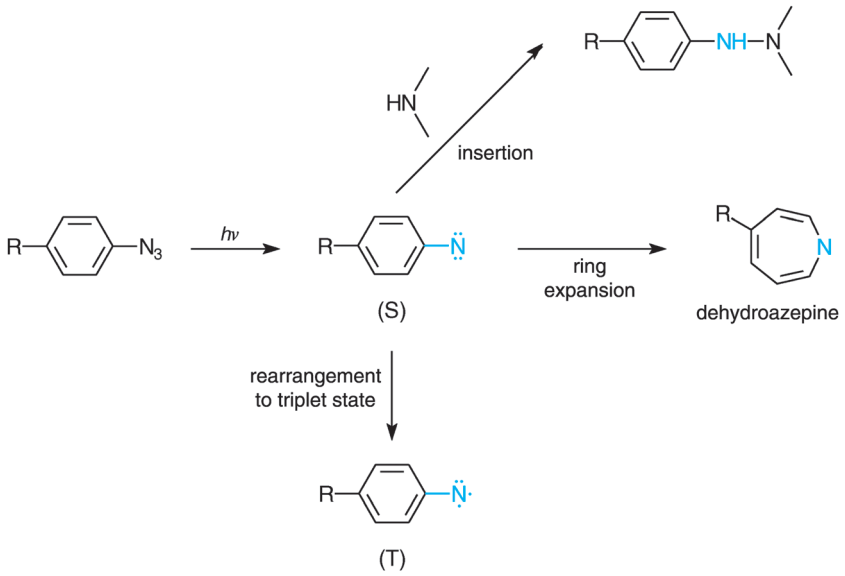


Figure 6.26 Reaction of aryl azides during irradiation. The resulting singlet nitrenes (S) can be inserted into N—H bonds, or they can be rearranged by ring expansion to a dehydroazepine.

Chemistry of Individual Reagents

Aryl Azides These compounds yield reactive nitrenes upon irradiation. Transiently formed singlet nitrenes (S in Figure 6.26) insert preferentially into N–H and O—H bonds or into other nucleophiles. Triplet nitrenes, on the other hand, can extract protons from C—H bonds. Aryl nitrenes undergo a rapid ring extension to form dehydroazepines. These are less reactive than nitrenes and react preferentially with nucleophiles and the solvent (e.g., water). In addition, dehydroazepines can react with the aryl azide to form a mixture of polymers. In the absence of nucleophiles in the ligand binding pocket of the protein the photolabeling fails.

Some of the aryl azides used for photolabeling of proteins are depicted in Table 6.4. The aryl fluoride (entry 1, Table 6.4) is relatively unspecific, reacting with amines, thiols, phenols, and other nucleophiles. *N*-Hydroxysuccinimides (entry 2), on the other hand, are applied frequently for photolabeling of amino groups. The ethyl amino group (3) allows photolabeling of proteins via their carboxy groups in the presence of carbodiimides.

The recently introduced perfluoro aryl azides (4) have the advantage of forming nitrenes with a far lower tendency to undergo ring extensions. Therefore, they react with a higher efficiency with C—H bonds.

With derivatives of azido salicylic acid (entries 5 and 6) radioactive iodine can be introduced directly into the photoactivatable compound (Figure 6.27). The sulfo group in (6) increases the water solubility considerably.

The esters (5) and (6) allow the introduction of groups to be iodinated into amino groups while iodoacetamide (7) is used mainly for the labeling of sulfhydryl groups. The presence of radioactivity in the label itself instead of a tyrosine residue favors the localization of the crosslink considerably. This is essential indeed if a cleavable crosslinker is used (see below). Aryl azides, especially those that carry the iodine in the same aromatic ring as the azide group, release iodine during photolysis. Therefore, mild photolysis conditions should be chosen. One

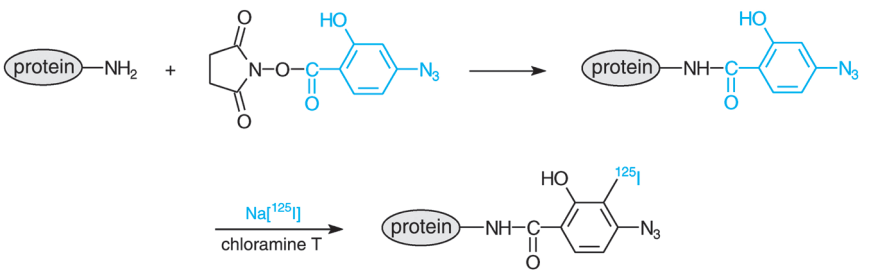


Figure 6.27 Modification with *N*-hydroxysuccinimidyl-4-azidosalicylic acid and subsequent radioactive labeling with chloramine-T.

should look for a compromise between preservation of the iodine and complete photoconversion of the photolabel.

Compounds shown in entries 8–10 are typical cleavable photolabeling reagents. Ester (9) can be labeled with radioactive iodine before using it for protein modification. After binding to protein amino groups the complex is irradiated. Subsequently the crosslink is cleaved by means of ß-mercaptoethanol or dithiothreitol. The result of this procedure is that only part of the reagent remains covalently bound to the protein. This makes the analysis much easier. A similar result can be obtained with the Denny–Jaffe reagent (10) by cleavage of the azo group with sodium dithionite.

Carbene Forming Reagents The most efficient starting compounds for obtaining carbenes are 1-trifluoromethy-1-phenyldiazirines.

Like the aryl azides, diazirines should also be handled in the dark or under dimmed light.

Photolysis of aryl(trifluoromethyl)diazirines (A in Figure 6.28) yields carbenes (B) that insert efficiently into C—H bonds to form stable products (C). Adducts with O—H and N—H bonds (E) are not stable. They eliminate HF and rapidly form difluoromethyl ketones and the starting compound. Furthermore, photolysis of (A) is accompanied by conversion into the diazo isomer (D). Because of the electron attracting effect of the CF_3 group this is fairly stable and does not form unwarranted side products.

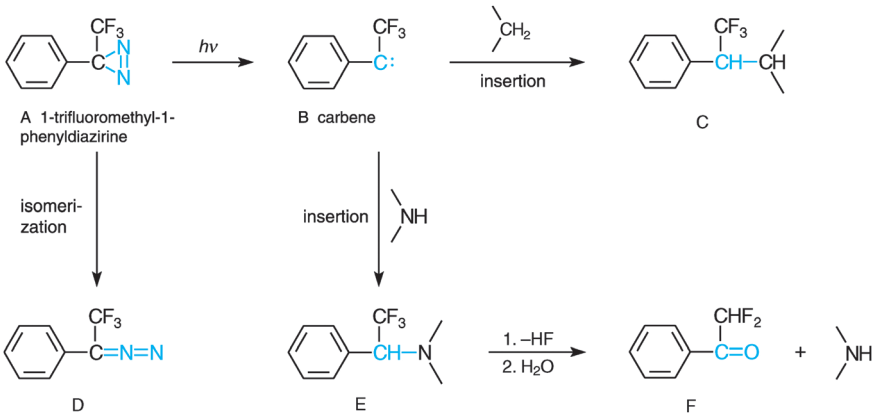


Figure 6.28 Carbene forming substances. 1-Trifluoromethyl-1-phenyldiazirine (A) forms after irradiation a carbene (B), which can insert into C—H and N—H bonds (E). The resulting alkyl compounds are stable while the trifluoromethyl amines eliminate HF and are hydrolyzed, resulting in trifluoromethyl ketones (F). The diazirine (A) also can rearrange to the corresponding diazo isomer (D).

Reagent 11 in Table 6.4 is used frequently for the nonspecific labeling of hydrophobic parts of membrane proteins, especially those that are in contact with lipids. The diazirine analogue of phenylalanine (entry 12) can be inserted into peptides in place of phenylalanine. The amino group specific reagent (13) is used to investigate binding pockets of ion channels and receptors. With reagent (14) an aryl-diazirinyl derivative can be attached via a thiol–disulfide exchange to a cysteine residue. After irradiation of the complex the S-S-bridge of the crosslink can be cleaved by means of a reagent agent.

The spacer arm in ester (15) can be cleaved by periodate oxidation. The advantage of this cleavable derivative is the small distance between the reactive groups. The diazirines (16) and (17) synthesized by Brunner contain an ester bond that can be cleaved by alkaline media. These reagents are available containing a very high specific radioactivity (about 2000 Ci $mmol^{-1}$). Photoactivation does not release iodine.

For crosslinking of specific nucleophilic groups of proteins diazopyruvoyl compounds (18 and 19 in Table 6.4) are chosen. *p*-Nitrophenyl diazopyruvate, for example (A in Figure 6.29), reacts with protein amino groups. Irradiation with UV light of 300 nm triggers a Wolff-rearrangement to ketenes (C). These can acylate nearby nucleophiles, forming intra- or intermolecular crosslinks (D). The advantage of diazopyruvoyl compounds is their stability towards thiols (which act by deactivation on aryl azides). Reagent (19) serves as a partner in reactions of diazopyruvoyl residues with –SH groups of proteins and peptides.

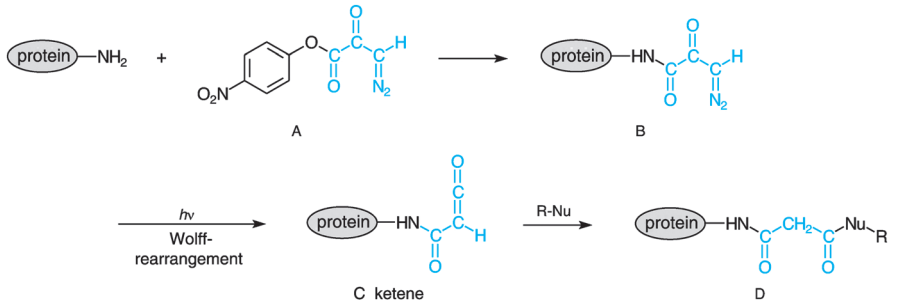


Figure 6.29 Reaction with *p*-nitrophenyl diazopyruvate (A). The reaction product (B) undergoes a Wolff-rearrangement to form the corresponding ketene (C), which subsequently forms with nucleophiles intra- or intermolecular crosslinks (D).

Benzophenone Photolabels As compared to diazirines or aryl azides, benzophenones are advantageous in several aspects: they are more stable and activation can be accomplished at longer wave lengths (350 nm). This reduces damage to the protein through irradiation. Benzophenones react preferentially with C—H bonds and the crosslinking yield is higher in most cases. On the other hand, sometimes the yield of photolabeling with an aryl azide was high but failed when the same ligand contained a benzophenone group. Apparently, the yield is increased by the hydrophobicity of the benzophenone on applying it to hydrophobic proteins, especially to membrane proteins.

Irradiation of benzophenones (Figure 6.30) results in the diradical triplet (B). Its interaction with C—H bonds causes the removal of hydrogen. Recombination of the intermediate ketyl (C) and alkyl (D) radicals yields the adduct (E) having a newly formed C—C bond. One characteristic of the triplet state originating from the benzophenones is their capability to return to the ground state. Therefore, they can be repeatedly undergo a cycle of excitation resulting in higher crosslinking yields.

The best results with benzophenones were obtained when phenylalanine residues in a peptide synthesis were replaced by Bpa (entry 20, Table 6.4). *N*-Hydroxysuccinimide (21) is used for the transfer of benzoylbenzoyl groups to protein amino groups. A tritium labeled hydroxysuccinimide is commercially available. 4-Maleimidobenzophenone (23) represents an example for a thiol specific photoactivatable benzophenone.

Finally, two methods should be mentioned that in principle can be applied with photo-reagents of different types. An interesting though not widely used version of photolabeling uses energy transfer. When a light sensitive group exists in a ligand–protein complex in close proximity with a tryptophan residue, photo-activation can take place via excitation of the tryptophan residue. The photolabel plays the role of an acceptor. Its absorption at the wavelength of the excitation of tryptophan should be low but should overlap considerably with its emission wavelength (320–340 nm). This procedure reduces non-specific reactions considerably.

Photoactivatable groups can be introduced into proteins and peptides not only by chemical modification or during peptide synthesis but also by means of biosynthesis. Because of various technical problems this method is not very frequently used. But, for example, in investigations of the transport of nascent proteins by ribosomes at the endoplasmic reticulum it was highly successful: in a first step lysyl-tRNA was converted with reagent 2 (Table 6.4) into *N*-(*p*-azidobenzoyl)lysyl-tRNA. The ribosomal machinery accepted this as a substrate like normal lysine-tRNA. The resulting protein contained several photoactivatable lysine residues, which might be a disadvantage. But this could be avoid by allowing for only one lysine codon in

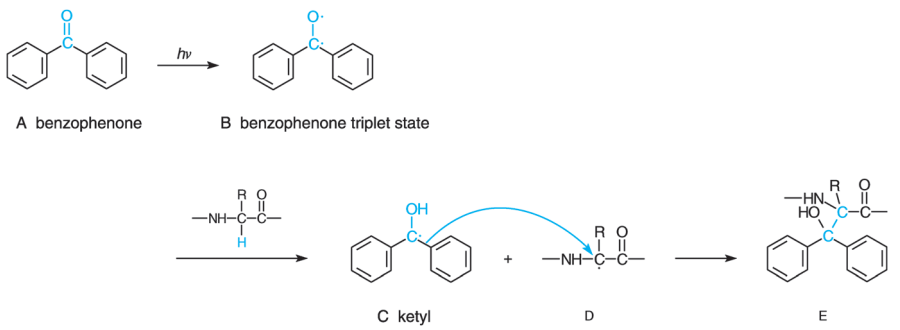
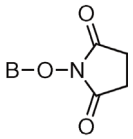
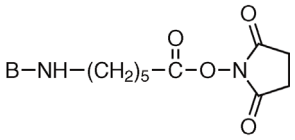
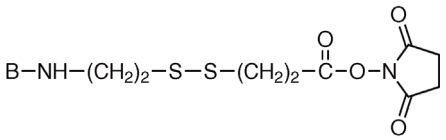
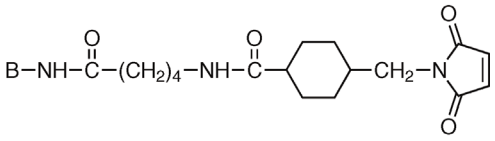
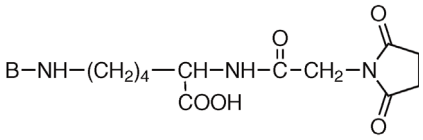
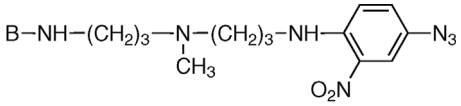
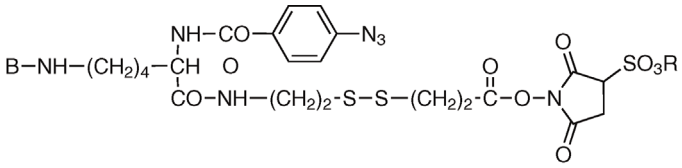
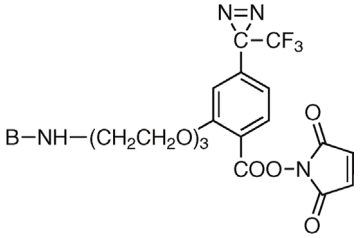
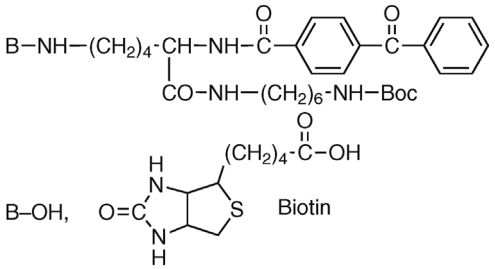


Figure 6.30 Photolabeling of peptides with benzophenone. Benzophenone (A) forms after irradiation a triplet state (B); the subsequently formed ketyl (C) and reacts with the resulting alkyl radical (D) to yield adduct (E) and a new C—C bond.

Table 6.5 Reagents for biotinylation (B: Biotin)

1		biotin *N*-hydroxysuccinimide ester
2		succinimidyl-6-(biotinamido)hexanoate
3		sulfosuccinimidyl-2-(biotinamido)-ethyl-1,3′-dithiopropionate
4		1-biotinamido-4-[4′-(maleimidomethyl)-cyclohexan-carboxamido]butan
5		3-(*N*-maleimidopropionyl)biocytin
6		photobiotin
7		2-[6-(biotinamido)-2-(*p*-azidobenzamido)-hexanamido]ethyl-1,3-dithiopropionate.
8		2-[2-[2-(2-biotin aminoethoxy)ethoxy]ethoxy]-4-[3-(trifluoromethyl)-3H-diazirin-3-yl]benzoic acid-*N*-hydroxysuccinimide ester
9		1-*N*[*N*$^{\alpha}$-(4-benzoyl benzoyl)-L-biocytinoyl]amino-6-(*N*-Boc-amino)hexan

the messenger RNA, with the others being removed by site-directed mutagenesis. Rapoport and coworkers distributed trifluoromethyldiazirinobenzoyllysine over the peptide preprolactin and labeled in this way the peptide transporting channel within the membrane of the endoplasmic reticulum.

Methods for Identifying the Site of a Crosslink As mentioned above, characterization of a crosslinked protein is easy if only the molecular weight of a labeled protein is to be determined. Gel chromatography and autoradiography may suffice. More detailed analysis should resort to mass spectrometry.

In most cases the chemistry of the photoreaction is not known precisely. In addition, the yields of the modification reaction are low. Therefore, the first hurdle is the isolation of the crosslinked peptides after enzymatic or chemical cleavage of the crosslinked complex. If mass spectrometry is applied, it is sufficient to identify the respective band on an SDS polyacrylamide gel. This can be cut out and processed as described in the Chapter 15 (Mass Spectrometry).

Some reagents, for example, the nitrodiazirines (compound 13 in Table 6.4) are chromophores, but identification at sub-nanomolar amounts requires a very high specific radioactivity. In any case, usually, many crosslinking products are obtained, which makes separation of the crosslinked peptides difficult.

Biotinyl groups can be added to photoactivatable molecules, thereby facilitating identification and isolation, for example, with an avidin matrix. Given the sensitivity of modern mass spectrometers, most classical methods for the isolation and enrichment of crosslinking products are nowadays no longer necessary and are omitted from this treatise. If available, antibodies can also be used for the isolation and characterization of crosslinks.

Edman Degradation, Section 14.1

Sometimes the instability of a crosslink may pose a problem. A good example of this is the crosslinking of hexaminidase: with a diazirine derivative of *N*-acetylgalactosamine, classical Edman sequencing of the isolated crosslinked peptides did not hint at any crosslinking. After treatment under alkaline conditions, at the respective position a glutamate instead of a glutamine residue was found. Obviously, the label was linked to the peptide through a labile ester bond.

A combination of Edman degradation with mass spectrometry (MALDI-MS) is the most reliable localization method. Many examples of this procedure can be found in the literature. Even crosslinked peptides in sub-picomolar amounts can be characterized today most reliably.

MALDI-MS, Section 15.1.1

Further Reading

Brunner, J. (1993) New photolabeling and crosslinking methods. *Annu. Rev. Biochem.*, **62**, 483–514.

Creighton, T.E. (1993) *Proteins: Structure and Molecular Properties*, W.H. Freeman & Co., New York.

Dorman, G. and Prestwich, G.D. (1994) Benzophenone photophores in biochemistry. *Biochemistry*, **33**, 5661–5673.

Haugland, R.P. (1996) *Handbook of Fluorescent Probes and Research Chemicals*, Molecular Probes, Inc.

Hermanson, G.T. (1996) *Bioconjugate Techniques*, Academic Press.

Hubbel, W.L. and Allenbach, C. (1994) Investigation of structure and dynamics in membrane proteins using site-directed spin labeling. *Curr. Opin. Struct. Biol.*, **4**, 566–573.

Imoto, T. and Yamada, H. (1989) Chemical modification, in Protein Function: A Practical Approach (ed T. E. Creighton;), IRL Press, pp. 247–277.

Kyte, J. (1995) *Structure in Protein Chemistry*, Garland Publishing, Inc., New York.

Means, G.E. and Feeney, R.E. (1971) *Chemical Modification of Proteins*, Holden-Day.

Rademann, J. (2004) Organic protein chemistry: drug discovery through chemical modification of proteins. *Angew. Chem. Int. Ed.*, **43**, 4554–4557.

Schuster, D.I., Probst, W.C., Ehrlich, G.K., and Singh, G. (1988) Photoaffinity labeling. *Photochem. Photobiol.*, **49**, 785–804.

Staros, J.V. (1988) Membrane-impermeant crosslinking reagents: probes of the structure and dynamics of membrane proteins. *Acc. Chem. Res.*, **21**, 435–441.

Tann, C.M., Qi, D., and Distefano, M.D. (2001) Enzyme design by chemical modification of protein scaffolds. *Curr. Opin. Chem. Biol.*, **5**, 696–704.

Wilchek, M. and Bayer, E.A. (eds) (1990) *Avidin–Biotin Technology*, Methods in Enzymology, vol. 184, Academic Press.

Zhang, J., Campbell, R.E., Ting, A.Y., and Tsien, R.Y. (2002) Creating new fluorescent probes for cell biology. *Nat. Rev. Mol. Cell Biol.*, **3**, 906–918.

Spectroscopy

7

Werner Mäntele

Johann Wolfgang Goethe-Universität Frankfurt am Main, Institut für Biophysik, Max von Laue Straße 1, 60438 Frankfurt am Main, Germany

Biophysical methods can help to obtain information on the form, size, structure, charges, molecular mass, function and dynamics, and other aspects of biological macromolecules.

Optical spectroscopic techniques can elucidate partial aspects of these questions – not in a highly detailed form as with the X-ray crystallographic picture of an enzyme, but with much less effort – concerning not only the technical complexity but also the requirements for purity and quantity of the samples and for special sample forms. In addition, typically, for spectroscopy less effort is needed to achieve a meaningful interpretation of the data, that is, data evaluation, model building, and experiments on model systems.

In relation to the information obtained, optical spectroscopic techniques require comparatively little effort. They have thus evolved to routine methods available in many laboratories, among them absorption spectroscopy in the ultraviolet and visible spectral range or fluorescence techniques. Absorption spectroscopy can be used for a quick determination of the concentration or purity of a protein preparation or for the analysis of pigments (e.g., in pigment–protein complexes).

Pigment A dye, that is, a molecule that absorbs light in the visible part of the electromagnetic spectrum and thus appears colored, for example, the carotins in carrots, the visual pigment rhodopsin in the retina of vertebrates, or the chlorophylls in photosynthesis.

Absorption methods are further used in the infrared spectral region, which is of increasing importance for the analysis of structure and function of biopolymers. Other methods rely on emission processes or on light scattering. Techniques that analyze light emission after absorption, such as fluorescence, have become important for the analysis of distances between functional groups in biopolymers or of folding, unfolding, and misfolding of proteins.

In contrast to the spectra of small molecules, which are, in general, well understood theoretically and can be modeled to fit the measured spectra, the spectral properties of biological macromolecules can sometimes only be described in a phenomenological way. Model calculations for spectra of these macromolecules frequently require complex normalization procedures to fit the observed spectra. Nevertheless, even empirical procedures in spectroscopy can help to identify and quantify molecules, to study their reactions, and to investigate their function in complex systems such as a living cell or a living organism. In some cases, they can be used on humans *in vivo* and have led to important applications in medicine.

This chapter starts with the physical principles of the interaction of light with matter as the basis of all optical spectroscopic techniques. Later, individual methods are discussed. The examples provided include established routine spectroscopic techniques in bioanalytical and biomedical spectroscopy as well as more recent applications from spectroscopic research.

The question of function and dynamics of biopolymers is closely related to the possibility of time-resolved spectroscopic measurements. We will thus discuss in connection with the individual methods more specific applications for different time domains, from seconds to femtoseconds.

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

7.1 Physical Principles and Measuring Techniques

7.1.1 Physical Principles of Optical Spectroscopic Techniques

All optical spectroscopic techniques rely on the same basic principle: electromagnetic radiation of a certain wavelength and intensity irradiates a sample and is either absorbed, scattered, or re-emitted after absorption (Figure 7.1).

In a photometric experiment, the electromagnetic radiation leaving the sample thus needs to be analyzed for its *intensity, wavelength*, and *angular distribution* (relative to the direction of the incoming radiation). The parameters for the incoming and the outgoing radiation can be used to calculate the sample properties.

Electromagnetic waves consist of an electric and a magnetic component that both depend on the position x and on the time t. This dependence is described by the oscillation of the electric ($\boldsymbol{E}$) and magnetic field vector ($\boldsymbol{H}$) in time and space:

$$\boldsymbol{E} = \boldsymbol{E}_0 \cdot \cos\left[2\pi\left(\nu t - \frac{x}{\lambda}\right) + \varphi\right] \quad \boldsymbol{H} = \boldsymbol{H}_0 \cdot \cos\left[2\pi\left(\nu t - \frac{x}{\lambda}\right) + \varphi\right] \tag{7.1}$$

Here, ν denotes the frequency of the light and λ the wavelength; φ is the phase angle for $x = 0$, $t = 0$. The wavelength λ and the frequency ν are related: $c = \nu\lambda$, where c is the speed of light. The speed of light depends on the medium, in vacuum it is $c_0 = 2.9979 \times 10^8\ \mathrm{m\,s^{-1}}$.

The electric field vector and the magnetic field vector are perpendicular to each other and to the direction of propagation. Electromagnetic waves are thus transversal waves. The direction of orientation for the electric (and the magnetic) field vector around the axis of propagation is termed the polarization of the light. *Natural light* is unpolarized; it shows an equal distribution of all directions of orientation. For *linear polarized* light, the electric field vector points in only one direction (with the magnetic field vector perpendicular to it), and for circular polarized light the field vector describes a spiral track around the direction of propagation.

The principle of a photometric experiment shown in Figure 7.1 can be extended using linear polarized light and some orientation of the "object". The measured intensity of the transmitted wave as a function of the direction of polarization can be related to the angle between the microscopic molecular orientation of the sample and the macroscopic orientation of the sample axis and the direction of polarization (Figure 7.2).

Polarized Light, as an electromagnetic wave, consists of an electric field component and a magnetic field component, each represented by a vector representing field strength and direction. The electric and magnetic field vector are oriented perpendicular to each other and perpendicular to the direction of propagation of light (Figure 7.2). The orientation of the electric or magnetic field vector is called the *polarization* of the light. For natural light, all directions are equally represented. If light is reflected from a surface or absorbed by oriented molecules, a specific direction of polarization can be favored; this is called *linear polarized light*. In spectroscopy, linear polarized light and circular polarized light can be used. For example, the linear polarization of sunlight from reflected surfaces is the basis on which honeybees orientate themselves, while circular polarized light is used for the function of liquid crystal displays.

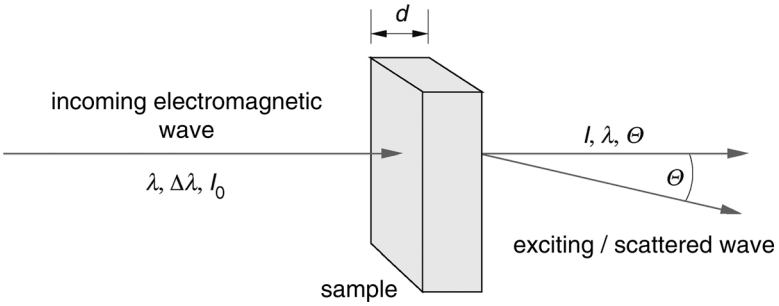


Figure 7.1 Principle of a photometric measurement. The incoming electromagnetic wave is characterized by the wavelength λ, the spectral bandwidth $\Delta\lambda$, and the intensity I_0. After passing through the sample for a path length d, the wave exits with intensity I, depending on the properties and the concentration of the absorber in the sample. In addition, the wavelength may be shifted by interaction with the sample, or scattering may lead to a deviation Θ from the original direction.

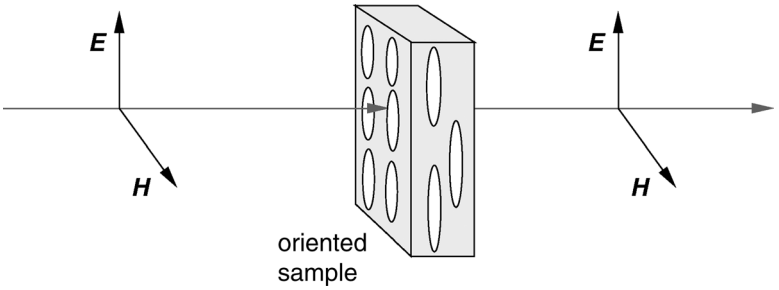


Figure 7.2 Principle of a photometric measurement with linear polarized light; $\boldsymbol{E}$ and $\boldsymbol{H}$ are the electric and magnetic field vectors, respectively.

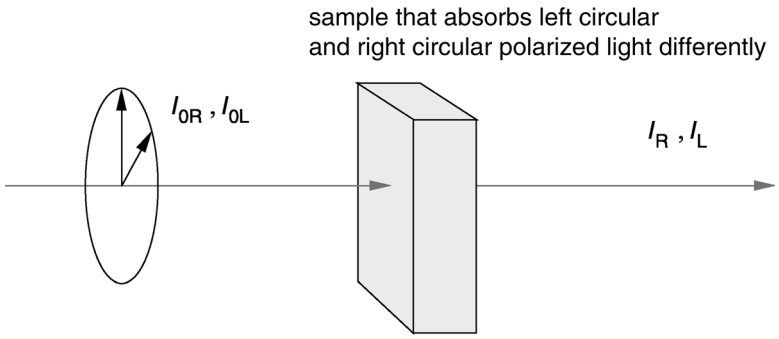


Figure 7.3 Principle of a photometric measurement with circular polarized light. I_{0R}, I_{0L}: Intensity of the right and left circular polarized component, respectively, of the incoming wave; I_R, I_L: intensity of the right and left circular polarized component, respectively, of the wave after passing through the sample.

If circular polarized light is used instead of linear polarized light, optically active samples can be studied. Instead of the macroscopic ordering of samples, this technique uses the properties of optically active samples that absorb left circular and right circular polarized light differently (Figure 7.3).

7.1.2 Interaction of Light with Matter

Electromagnetic radiation can be described as waves or as a steady flow of particles; the synthesis of both descriptions is called wave-particle dualism. The particle description implies a steady flow of zero-mass particles termed photons that move at the speed of light, each photon having the energy $E = h\nu$. Here, h is Planck's quantum of action, 6.62×10^{-34} J s, and ν is the frequency of light in s^{-1}.

Arguments for the use of the particle nature of light are, for example, the quantized response of the photoreceptors in the eye or the properties of technical photodetectors. In the description of light as a wave it is discussed as a transversal electromagnetic wave with the properties discussed above. Arguments for the wave nature of light come from phenomena such as refraction, diffraction, or interference of light. If the particle and the wave description are combined, one can use terms like "photon energies" and at the same time use the term "wavelength". A phrase like ". . . a photon of the wavelength 500 nm" illustrates this dualism.

The basics of spectroscopy discussed in this chapter imply photons in the ultraviolet (UV), visible (VIS), near-infrared (NIR), and the mid-infrared (MIR) spectral region. At lower photon energies – or longer wavelengths – the spectral range may be called the far-infrared (FIR) or terahertz (THz), and at even lower photon energies the microwave and radio wave ranges follow, both of which are particularly relevant for magnetic resonance techniques such as EPR and NMR.

EPR, Section 18.2

NMR, Section 18.1

Table 7.1 additionally lists the "wavenumbers" (in the unit cm^{-1}) for all spectral regions mentioned. The unit "wavenumber" corresponds to the number of waves per cm and is preferred by spectroscopists, in particular for infrared and Raman spectroscopy. One advantage of using wavenumber units is that numbers are proportional to the energy of the photons used.

The unit Joule is not very convenient for use with photon energies, because it is derived from macroscopic units such as meter, kilogram, and second. The energy of a photon may be obtained from the relation $E = h\nu$, but the numbers obtained are very small and require handling decimals. Table 7.1 thus additionally lists the energy in electron volts (eV), which gives handy numbers, in the fifth column. Photon energies given in electron volts (eV) are particularly useful if electron transfer reactions following the absorption of photons are analyzed, as is the case for the primary reactions in photosynthesis.

If we consider that the binding energies between atoms in a molecule are up to a few electron volts, it is evident from Table 7.1 that the absorption of electromagnetic radiation in bioanalytics does not break bonds nor does it ionize molecules (with the exception of radiation in the X-ray domain, in the ultraviolet C, and eventually in the ultraviolet B region). The typical energy required to break bonds or to ionize atoms or molecules is a few eV. However, the absorption of a photon may result in the transition of an electron to a higher orbital. This process is termed an electronic transition.

The unit electron volt is defined as follows: An electron in a vacuum tube emerging from a cathode is accelerated towards an anode. If the acceleration voltage between cathode and anode amounts to 1 volt, the electron draws from this electric field the energy 1 eV. It is manifested in its kinetic energy E_{kin} upon reaching the anode; E_{kin} is given by the product of the elementary charge e (in Coulomb) (1.6×10^{-19} C) and the potential difference between cathode and anode (1 volt): $1\,eV = 1.6 \times 10^{-19}\,C\,V = 1.6 \times 10^{-19}\,A\,sV = 1.6 \times 10^{-19}\,W\,s = 1.6 \cdot 10^{-19}\,J$.

Table 7.1 Spectral range for different types of spectroscopy (UV: ultraviolet; VIS: visible light; NIR: near-infrared; MIR: mid-infrared; FIR: far-infrared; THz: terahertz).

Range	Wavelength	Wavenumber (cm^{-1})	Photon energy (J)	Photon energy (eV)	Excitation/Interaction	Applications
X-rays	0.01–100 nm	10^9–10^5	2×10^{-14}–2×10^{-18}	10^5–12.4	Inner electrons	Structural studies
UV-C	100–280 nm	100 000–35 000	2×10^{-18}–7×10^{-19}	12.4–4.4	Outer electrons	UV/VIS-Spectroscopy
UV-B	280–320 nm	35 000–30 000	7×10^{-19}–6.25×10^{-19}	4.4–3.9	Delocalized electrons	
UV-A	320–400 nm	30 000–25 000	6.25×10^{-19}–5×10^{-19}	3.9–3.1	UV proteins/pigments	
VIS	400–760 nm	25 000–13 000	5×10^{-19}–2.6×10^{-19}	3.1–1.6	Pigments	
NIR	760–3000 nm	13 000–3300	2.6×10^{-19}–6.6×10^{-20}	1.6–0.4	Overtones	Industry: QC
MIR	3–30 μm	3300–330	6.6×10^{-20}–6.6×10^{-21}	0.4–0.04	Vibrational levels	IR/Raman
FIR/THz	30–1000 μm	330–10	6.6×10^{-21}–2×10^{-22}	0.04–0.001	Rotational levels	THz imaging
Microwaves	1000 μm–10 mm	10–0.1	2×10^{-22}–2×10^{-24}	10^{-3}–10^{-5}	Electron spins	EPR
Radio waves	10 cm–100 m	10^{-1}–10^{-4}	2×10^{-24}–2×10^{-27}	10^{-5}–10^{-8}	Nuclear spins	NMR

The description of light as an electromagnetic wave is typically used to discuss the interaction of light with matter. For the first step of this discussion, we focus on the electric field vector $\boldsymbol{E}$ since its properties dominate the interaction over wide spectral ranges. The magnetic field vector $\boldsymbol{H}$ is relevant for resonance methods. Strictly speaking, a discussion of the interaction of light with matter should include the temporal and the spatial variation of the electric field vector $\boldsymbol{E}(\boldsymbol{x},\boldsymbol{t})$. However, the dimensions of the absorbing molecules are typically very small as compared to the wavelength λ that is absorbed by them. If we take as an example the molecule retinal (shown below in Figure 7.11) with a maximum lateral dimension of 15–20 Å, its absorbance is at around 360 nm (or 3600 Å) in the isolated form. Thus, the electric field vector (and the magnetic field vector, too) across the molecule can be considered as spatially constant, as a good approximation.

With this (well justified) approximation, we just need to consider the time dependence of the electric field vector in the molecule to describe its interaction with matter. It can be described as follows:

$$\vec{\boldsymbol{E}} = \vec{E}(t) = \vec{E}_0 \cos(\omega t) = \vec{E}_0 \cos(2\pi\nu t) = \vec{E}_0 e^{i\omega t} \tag{7.2}$$

where: $\vec{E}_0$ denotes the maximum value of the electric field,
ω is the angular frequency,
i is the imaginary number ($\sqrt{-1}$),
ν is the frequency.

Properties of Molecules To discuss the interaction of the electric field vector with a molecule we first consider the total energy of the molecule, which is composed of movements of the atomic nuclei and the electrons. In a description using classical mechanics, the nuclei exhibit translational, vibrational, and rotational movements, and electrons follow nuclei in discrete energy levels. In a more detailed, but still classical, description electrons are found in shells and orbitals that are occupied according to predetermined rules. Quantum mechanically, these orbitals are described by space- and time-dependent wave functions. The square of a wave function is related to the probability of finding an electron at a given interval of space and time.

To describe molecules composed from atoms, this orbital model is extended from atomic orbitals (AOs) to molecular orbitals (MOs), for example, by forming linear combinations of atomic orbitals (LCAO) to molecular orbitals. Box 7.1 summarizes some definitions used for the description of orbitals.

For simple molecules (e.g., C_2H_4, ethene) the basic types of molecular orbitals can be directly given for the ground and the excited state. For the bonding σ orbital the electron density is localized along a line connecting both carbon nuclei. Its counterpart in the excited state, the σ* orbital, exhibits zero electron density in the region between the carbon nuclei. It is thus called antibonding, because it does not contribute to the compensation of the nuclei repulsion. A

Box 7.1 Glossary of terms and rules used in atomic and molecular physics.

- **Spin S:** If the electron is considered as a spatial distribution of negative charge, its rotation with a given angular momentum yields a magnetic moment.
- Each **orbital** can only be occupied by two electrons. The **Pauli Principle** (according to Wolfgang Pauli) states that their spin has to be antiparallel. This is usually represented as {+ ½; −½} or {−½; +½}.
- **Multiplicity** is defined as the sum of the magnetic moments in an orbital calculated by the rule $\boldsymbol{M = 2\,|S| + 1}$. Accordingly, the multiplicity of a spin configuration {+ ½; −½} or {−½; +½} yields $S = 0$ and thus $M = 1$; this is termed a **singlet state**. It may happen that in a photochemical process a spin is inverted to yield spin configuration {+½; +½} or {−½; −½}, which in turn yields $S = \pm 1$ and thus $M = 3$; this is termed a **triplet state**.

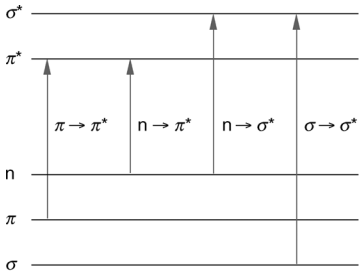


Figure 7.4 Molecular orbital levels and possible electronic transitions between them. The respective length of the arrows corresponds to the energy of the photons required to excite the transition.

bonding π orbital exhibits maximum electron density in a plane perpendicular to the axis connecting the nuclei, but has minimal electron density at this axis (a knot). Its counterpart in the excited state, the antibonding π^* orbital, does not show a maximum of electron density between the nuclei, in analogy to the σ^* orbital. A double bond implies four electrons, two in a σ orbital and two in a π orbital. When there are π orbitals that are not involved in binding, they are termed non-bonding (n). Figure 7.4 shows possible transitions between these orbitals.

Electrons exhibit an angular momentum and thus a magnetic moment, termed spin (see also glossary in Box 7.1). A stable molecule will typically exhibit molecular orbitals containing two electrons with antiparallel spin, as described by the Pauli principle.

Absorption of a Photon Starting from the given configuration of nuclei and electrons in a molecule (sometimes termed the ground state, but this will be detailed later), absorption of a photon can lead to transition of an electron from an occupied into an empty molecular orbital. In the terms used for photochemistry, this is indicated by designating the initial orbital and the final orbital.

A $\pi \rightarrow \pi^*$ transition in a molecule occurs if an electron is excited from a π orbital to a π^* orbital. The transition to a previously empty orbital may occur with the conservation of the magnetic moment (the spin) of the electron, but can also involve spin inversion.

The electric field vector of an incoming electromagnetic wave can induce dipoles in the charge distribution of the molecular structure. This induced dipole, μ_{ind}, is proportional to the electric field strength $\boldsymbol{E}$:

$$\boldsymbol{\mu}_{ind} = \alpha \boldsymbol{E} \tag{7.3}$$

In this equation, the electric field is represented as a vector. Consequently, the dipole moment induced in the molecule is vectorial, too, and exhibits a clear orientation within the geometry of the molecule. The constant α, called the polarizability of the molecule, depends on the electronic structure of the molecule and may differ in all three directions of space. It must thus be written as a tensor. A somewhat descriptive interpretation of a high polarizability is that the electronic system of the molecule may be easily "deformed" by an external electric field.

Because of the variation of the electric field E with time (Equation 7.2), the induced dipole moment also varies with the angular frequency ω. The magnitude of the induced dipole μ_{ind}, controlled by the polarizability α, determines the probability for the transition of the molecule to a different state upon interaction with the electromagnetic wave.

In molecular spectroscopy, the term transition dipole moment denotes the shift of electric charges in the molecule upon the transition between two electronic states. Historically, transitions are classified as "allowed" if they are highly probable, and as "forbidden" if they are highly improbable. For an allowed transition corresponding to strong absorbance of the substance, the transition dipole moment can correspond to the shift of an elementary charge (1.6×10^{-19} C) by the length of a chemical bond (approx. 10^{-10} m). The probability of a transition, which is reflected in the intensity of the absorption or emission, depends on the square of the transition dipole moment.

Energy Levels of Molecules To describe the transitions between different energy levels in molecules, let us first define the individual contributions to the total energy of the molecule. These contributions arise from the energy of the electronic states and from the movements of the

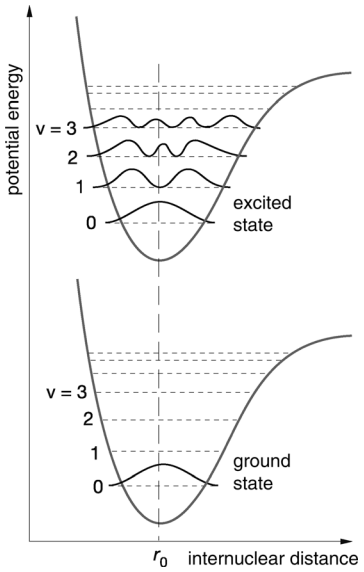


Figure 7.5 Schematic representation of the energy levels for a diatomic molecule. The potential curves for the two electronic levels S_0 and S_1 are shown with the vibrational levels $v = 1, 2, 3$, and so on; r_0 is the equilibrium distance. The curves at the different vibrational levels represent the probabilities of finding the nucleus in this state.

nuclei as well as from the energy due to magnetic properties of both. These different energy contributions each differ by 1–2 orders in magnitude:

$$E_{tot} = E_{el} + E_{vib} + E_{rot} + E_{magn} \tag{7.4}$$

where E_{tot} is the total energy of the molecule. It is composed of the energy corresponding to the electronic levels E_{el}, of the vibrational energy of the nuclei E_{vib} (they can vibrate relative to each other), of the rotational energies of atoms or atom groups around a common axis E_{rot}, and of the magnetic energies of the nuclei and the electron orbitals, E_{magn}. For the discussion of optical spectroscopy, we will start by neglecting these magnetic properties.

It is already quite complicated to visualize these energy contributions as a function of atomic coordinates for a small molecule consisting of several atoms, and not feasible for a polyatomic biopolymer, at least not as a whole. Thus, a diatomic "model" molecule is typically chosen, where a molecular parameter, for example, the distance of the nuclei, is drawn on the abscissa and the potential energy of the molecule is drawn on the ordinate. In the electronic ground state S_0 (for a definition see below), a potential curve is obtained with a minimum that corresponds to the equilibrium distance of the nuclei, r_0. In a classical description, one might describe this molecule as two masses (the two atoms) connected by a spring (the chemical bond). Figure 7.5 shows how these potential curves are associated with two different electronic levels.

For small oscillations around the equilibrium distance r_0 the potential curve can be reasonably approximated by a parabola, often denoted as the *parabolic potential* or *harmonic potential*. The vibrations thus calculated classically or quantum mechanically are that of a harmonic oscillator. As the distance between the nuclei is reduced, the energy increases more strongly than for the parabolic potential due to the electrostatic repulsion of the positively charged nuclei. As the distance of the nuclei is increased, the parabolic potential becomes increasingly anharmonic and develops to an energy limit that corresponds to the energy of the dissociated atoms, the so-called dissociation limit. An analogous potential curve can be drawn for the first electronically excited. It is positioned at higher energy than the ground state. Furthermore, its equilibrium distance r_0 may be different due to different bonding forces in the altered electronic configuration.

Within these two potential curves S_0 and S_1 are vibrational levels v_1, v_2, v_3, up to the dissociation limit, that is, to an energy where the atoms are so far apart that the bond can break. For small vibrational amplitudes around r_0, the parabolic potential can be used as an approximation, and the energy levels of the vibrations are at equal distance, given by Equation 7.5:

$$E_v = h\nu\left(v + \tfrac{1}{2}\right) \tag{7.5}$$

In this equation, ν denotes the vibrational frequency of the classical harmonic oscillator, h is Planck's quantum of action, and v is the vibrational quantum number, which denotes the vibrational level. Please note that for $v = 0$ the energy is equal to $E_0 = ½h\upsilon$ and is thus >0. This minimal energy is denoted as the zero-point-energy and is a consequence of the Heisenberg uncertainty principle, which states that the spatial coordinates and the energy cannot be given simultaneously at arbitrary precision (which would be the case if the oscillator would be at the equilibrium position with a fixed energy and distance).

For wider displacements from the equilibrium position, the vibration becomes increasingly anharmonic, and the approximation of a harmonic oscillator is no longer valid. This anharmonicity leads to a decrease of the distances between the energy levels and finally to a continuum at the dissociation limit.

Molecules show vibrational and rotational movements that may be coupled. Consequently, the vibrational levels are subdivided into rotational levels $r_1, r_2, r_3 \ldots r_n$. We will discuss the vibration forms of small molecules in Section 7.3.

A simple estimation shows that, at room temperature, all molecules are in the electronic ground state S_0. The energy gap between the ground state S_0 and the first excited state S_1 is typically at least 1 eV (this corresponds to the energy of a photon at approx. 1200 nm). The energy gap between the lowest vibrational level ($v = 0$) and the next higher vibrational level ($v = 1$) is typically 0.1–0.5 eV (which corresponds to the energy of a photon in the mid-infrared, see Table 7.1). The thermal energy at room temperature is given by $E_{therm} = kT$, where k is the Boltzmann constant and T the absolute temperature in Kelvin. E_{therm} at room

temperature is approx. 0.025 eV, which is not sufficient for electronic or vibronic excitation. We thus conclude that, in thermal equilibrium at room temperature, all molecules are in the electronic ground state, and almost all molecules are in the vibrational ground state. Only rotational levels are significantly occupied. This molecular configuration is denoted as (S_0, $v = 0$, r_1 . . . r_n).

Transitions between Energy Levels Starting from the state (S_0, $v = 0$, r_1 . . . r_n), absorption of a photon with a suitable energy can lead to transitions between different levels. Absorption of a photon in the IR excites rotational and vibrational transitions within S_0. Absorption of a photon in the UV, the visible, or the near-infrared range can excite electronic transitions between S_0 and S_1 (or higher excited states).

The condition for such a transition to happen is that the energy difference between the initial and the final state corresponds to the energy of the photon. However, not all transitions in this scheme are equally probable. There are a number of selection rules that render some transitions "allowed" (i.e., highly probable and thus related with strong absorption) or "forbidden" (i.e., improbable and thus related with weak absorption).

One of these selection rules is known as the Franck–Condon principle, displayed in Figure 7.5. It is based on the fact that the equilibrium distance of the nuclei in the electronically excited state may be either equal, or smaller, or larger than in the ground state. Only the first case is displayed in Figure 7.5. It is further based on the fact that the electronic transition from S_0 to S_1 proceeds rapidly (in the order of $<10^{-15}$ s), much faster than the vibrations of the nuclei with respect to each other (one vibrational period takes about 10^{-13} s). Thus, the nuclear distance remains almost constant during an electronic transition. This is represented in Figure 7.5 by a vertical line representing the electronic transition. This line leads to an excited vibrational level whenever the equilibrium distance in the excited state deviates from that in the ground state. Electronic and vibrational ("vibronic") transitions are thus coupled.

In the quantum mechanical description of this phenomenon, the overlap of the wave functions in the ground state S_0 and the excited states S_1. . . . S_n is considered. These wave functions represent the probability that a state is occupied. In the lowest vibrational state ($v = 0$), the wave function exhibits a maximum for the equilibrium distance r_0, for all higher vibrational levels this maximum is at the turning points (see Figure 7.5). Because the electronic transitions are considerably faster than the vibrations of the nuclei, the most probable transitions are those where the vibrational wave functions (and thus the probabilities for being occupied) in the ground state and in the excited states show highest similarity and overlap.

As discussed above, an electronic transition typically occurs with the spin conserved. In this case, the spin in the initial state is described as $S = +½/-½$, which is identical to that of the final state: $S = +½/-½$. The multiplicity in both cases is $M = 1$; the initial and final states are thus termed singlet states, and the transition is called a singlet transition. However, there are conditions under which a spin changes its direction during a transition, it "inverts." This leads from a spin configuration $S = +½/-½$ to a configuration $S = +½/+½$ or $S = -½/-½$, that is, from a singlet state to a triplet state. This transition is called intersystem crossing. The "intercombination rule" states that transitions between singlet and triplet states, that is, transitions with spin inversion, are normally forbidden. Practically, this means that they are improbable and thus related with only weak absorption. However, they are not exotic and they are important for some photochemical processes in biological systems.

The general condition for a transition is the matching of the energy difference between the initial and the final state with the energy of the incoming photon. However, not all transitions in this scheme are equally probable. There are several selection rules that govern the probability of transitions and lead to "allowed transitions", that is, transitions that are highly probable and thus related to strong absorption, and to "forbidden transitions", that is, transitions that are improbable and thus related to weak absorbance.

The Jablonski Diagram Energy levels and transitions in polyatomic molecules cannot be displayed as easily with potential functions as they are for the diatomic molecule in Figure 7.5. Instead, a simplified term scheme, also called a Jablonski diagram, is used to describe these processes independent of the size or complexity of the molecule. Figure 7.6 shows a term scheme with the electronic ground state (S_0) and a first and second electronically excited state, S_1 and S_2. In addition, the first and second triplet states (T_1 and T_2) are shown. The individual electronic states are further subdivided into vibrational levels (v_0, v_1, v_2, v_3 . . . v_n). The vibrational levels are further subdivided into rotational levels (r_0, r_1, r_2, r_3, . . . r_n); for a better overview, they are shown in the enlarged view on the right-hand side.

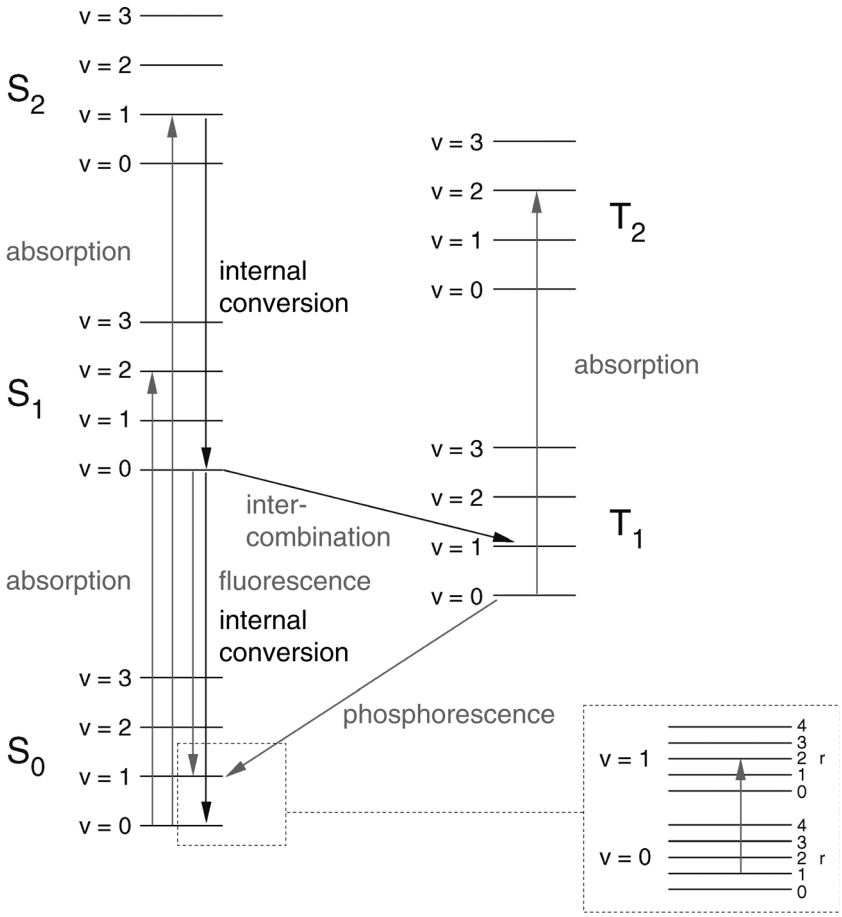


Figure 7.6 Term scheme of a molecule with possible electronic and vibronic transitions. In the enlarged view of $v = 0$ and $v = 1$, a vibrational transition with an energy corresponding to infrared radiation is shown. Radiationless transitions are shown in black. Transitions that are related to absorption or emission of a photon are shown in gray.

The transitions related to absorption or emission of a photon are denoted as follows:

- Singlet absorption: S_0, $v = 0 \rightarrow S_1$, $v = 2$, or S_0, $v = 0 \rightarrow S_2$, $v = 1$
- Triplet absorption: T_1, $v = 0 \rightarrow T_2$, $v = 2$
- Fluorescence: S_1, $v = 0 \rightarrow S_0$, $v = 1$
- Phosphorescence: T_1, $v = 0 \rightarrow S_0$, $v = 1$
- Infrared absorption: S_0, $v = 0 \rightarrow S_0$, $v = 1$

In addition, the radiationless transition (*internal conversion*) and the transition from the singlet to the triplet state (*intercombination*, *intersystem crossing*) are also shown. In the enlarged view, a transition corresponding to an infrared absorption $v = 0$ ($r = 1$) $\rightarrow v = 1$ ($r = 2$) is displayed.

For a better overview, the electronic and vibrational ("vibronic") levels of molecules are almost always drawn in the form of a term scheme (Figure 7.6). To fully understand the possible molecular transitions, it is important to note that higher excited vibrational levels of the electronic ground state S_0 can reach up to the energy of the first electronic excited state S_1, and correspondingly the higher excited vibrational levels of S_1 can reach up to the energy of the second electronic excited level S_2. Electronic and vibrational levels are thus closely coupled. This explains how electronically excited states can also deactivate in a radiationless form, a process that is denoted as *inner conversion*.

After excitation, for example, from S_0, $v = 0$ to S_1, $v = 2$, the molecule is for a very short time in the higher excited vibrational level, and not in thermal equilibrium with its environment. In this case, vibrational levels are occupied that would otherwise only be reached at higher temperature. It is acceptable to call this a "hot" molecule, which will return rapidly (in approx. 10^{-12}–10^{-13} s) to its thermal equilibrium. This process is sometimes called "thermalization"; the energy is dissipated to the molecule's environment, for example, to the solvent.

The time required to return from one of the excited states to a state of lower energy or to the ground state can be described by the "lifetime" of the state. A state depopulating by a transition with a high probability will show a short lifetime, while a state that deactivates by a "forbidden" transition will show higher lifetime. As an example, the return from the first excited singlet state (S_1) to the ground state (S_0) by fluorescence will show a high transition probability due to the fact that no spin inversion is involved. In this case, the lifetime is of the order of 10^{-9} s. For return from the first excited triplet state (T_1) by phosphorescence, spin inversion is required; this transition is much less probable, and the lifetime can range from milliseconds to seconds. This estimation of lifetimes, however, is much simplified, since radiationless processes can compete with light emitting processes. In this case, shorter lifetimes are observed since the excited state is depopulated faster by two or more parallel occurring reactions (much as a bucket empties faster with two or more holes in the bottom than with a single hole).

Using this term scheme, all absorption and emission processes of molecules relevant for bioanalytics can be reasonably described. It should be emphasized that this scheme is only a coarse approximation for the transitions. Other processes also contribute, for example, the radiationless energy transfer from an excited electronic state to a second, neighboring molecule. This energy transfer, if taking place from a singlet state and towards a singlet state, is termed singlet-singlet energy transfer.

Apart from "radiating" and "radiationless" transitions, photochemical reactions can occur, for example, the fast transfer of an electron to a neighboring molecule. This can lead to coupling of redox processes (an oxidation coupled with a reduction) to light-induced reactions. Reactions of this type play a vital role in the primary processes of photosynthesis.

If we consider a molecule in a "real" environment, that is, a molecule in solution or in a protein matrix, we have to extend the model used in the term scheme. The electron distribution in the excited state is much more extended in space than the ground state. This leads to a stronger interaction of the excited state with a polar solvent as compared to the ground state. For a $\pi \rightarrow \pi^*$ transition, the energy level is thus lowered and a redshift of the respective transition is observed. In contrast, an $n \rightarrow \pi^*$ transition shows a stronger interaction with the polar medium in the ground state; this leads to a blue-shift. In addition to modifying the energy levels, the interaction of chromophoric groups among each other and with the medium also can modify the transition dipole moments and thus change transition probabilities and absorption strengths.

Discussion of the physical effects that increase or decrease absorption dependent on the orientation of transition dipole moments with respect to each other is beyond the scope of this chapter. A straightforward phenomenological description of the absorption of a chromophore follows a simple terminology: a bathochromic shift corresponds to a redshift and a hypsochromic shift to a blue-shift of the absorption maximum, while an increased absorption strength is termed hyperchromism and a reduced absorption strength is known as hypochromism.

The term scheme of a molecule can now be easily "translated" into a spectrum that can be recorded with a suitable spectrophotometer. Absorption only takes place if the energy of the incoming photons $E = h\nu = hc/\lambda$ corresponds to the "energy gap" of the molecule, that is, the distance between S_0 and S_1 or S_0 and S_2, and so on. Consequently, the spectrum derived from this description would consist only of lines due to the discrete energies of the levels S_0, S_1, S_2, . . . S_n.

Even if the subdivision in vibrational levels and rotational levels is added, with the possibility that transitions occur from the lowest vibrational level of the electronic ground state S_0 to different higher vibrational levels of S_1 (S_1, $v = 1$, $v = 2$, $v = 3$, . . .), only line spectra would be observed, with a series of absorption lines of different strength.

Broadening of absorption lines is due to the finite lifetime of the excited states, to the Doppler effect (molecules may move towards the source or away from the source while emitting, leading to higher or lower frequencies; an example is the sound of police cars moving towards or away from the observer), or to collisions with other molecules. Practically, this is only the case for gases at low pressure; for molecules in solution these vibrationally-structured spectra are broadened to absorption bands.

An absorption band is characterized by its absorption maximum at a certain wavelength (λ), by its height (the absorption A), and its half-width, typically measured as full-width at half-maximum (FWHM).

Absorption bands of most biological molecules in their "real" environment exhibit half-widths that are mostly a consequence of vibrational substructures and heterogeneity in the local

conformation of the chromophore and its interaction with the environment. Broadening by a finite lifetime or by collisions is not relevant.

Because the vibrational substructures and the heterogeneous environment are the main source of broadening (this is frequently termed "heterogeneous broadening"), the half-widths of electronic transitions strongly depend on temperature. At low temperature, higher vibrational or rotational levels are less occupied, thus resulting in transitions with a lower half-width ("sharper transitions"). In addition, certain protein conformations are "frozen" at low temperature, which also leads to a reduction of half-widths. For the spectroscopy of biological molecules, these effects can be used with cryostats, that is, sample chambers to cool samples down to the temperature of liquid nitrogen (77 K) or even to that of liquid helium (4 K). This results in significantly sharper bands and allows us to separate different pigment populations that otherwise would hide within the broad band envelope. A well-known example of this procedure is the low-temperature spectroscopy of pigment populations in photosynthetic pigment–protein complexes.

7.1.3 Absorption Measurement and the Lambert–Beer Law

The Lambert–Beer law describes the relation between a macroscopic (in the laboratory) measurable quantity, the absorption of electromagnetic radiation with a given photon energy or wavelength, and the microscopic properties of an atom or molecule. To use this relation, the absorbing substance must be homogeneously distributed in solution and there must be no light scattering or photochemical reactions.

For monochromatic light, that is, photons of (ideally) a single precise energy, the absorption A is defined as:

$$A = \log\left(\frac{I_0}{I}\right) = \varepsilon c d \tag{7.6}$$

In Equation 7.6, I_0 and I denote the intensity of the radiation entering and leaving the analyzed solution, respectively, c is the concentration of the absorbing substance (mol l^{-1}), and d is the path length of the beam passing through the solution. The proportionality constant ε is called the molar absorption coefficient (occasionally also known as the "extinction coefficient", see below). Because the logarithm is dimensionless, the absorption is a dimensionless quantity, although one may find sometimes incorrect units such as "absorption units" (AU) or "optical density units" (OD). On a strict view, both are incorrect. The molecules in solution are typically analyzed in sample containers, so-called "cuvettes," with an optical path length in the mm or cm range. The path length is thus usually measured in cm. To become dimensionless, ε has the dimension $1\,\mathrm{mol}^{-1}\,\mathrm{cm}^{-1}$. Instead of the absorption A, the transmission T or the percent transmission $T_{\%}$ may be used:

$$T = \frac{I}{I_0} \qquad T_{\%} = 100\,\frac{I}{I_0} \tag{7.7}$$

The relation between absorption and transmission and between the light intensities I and I_0, which have to be measured precisely by a detector, indicates the meaningful range of an absorption measurement. For an absorption of 1, corresponding to a transmission of 10%, a photometer has to detect and process with high linearity and precision intensities that differ by a factor of 10. For an absorption of $A = 2$, the intensities I_0 and I differ by a factor of 100, and for $A = 3$ by a factor of 1000. Modern electronics and computerized digital photometers tend to suggest precise absorption measurements even for an absorption value of 3 or higher – something that instrument manufacturers like to advertise. However, if we consider the logarithmic relation of absorption and intensities, absorption values above 2 should not be taken too seriously. If a really precise absorption measurement is required, an absorption value of approx. 1 should be aimed for, either by using cuvettes of shorter or longer path length or by diluting or concentrating the sample solution.

As argued above, the Lambert–Beer law is only valid in a strict form for monochromatic light. The I, I_0, and ε, and with them also A and T, are thus wavelength dependent. It is quite common to denote this by giving the respective wavelength, for example, indexed by the wavelength λ. If the absorption or transmission is drawn as a function of the wavelength λ, an absorption spectrum or a transmission spectrum is obtained. A more professional form is to draw the absorption coefficient ε as a function of the wavelength; this is independent of experimental parameters such as path length or concentration.

If the analyzed sample contains several absorbing components, the absorption is additive, provided that the components are not interacting. For this case, the Lambert–Beer law can be extended:

$$A = (\varepsilon_1 c_1 + \varepsilon_2 c_2 + \ldots + \varepsilon_n c_n) d \tag{7.8}$$

The terms "absorption" and "extinction" are often used synonymously in text books; however, they should be used more precisely. Absorption is related directly to the molecular process of the interaction of light with matter by the Lambert–Beer law. Extinction (from the Latin *extingere*, that is, erase) denotes the fraction of light removed from the measuring beam by the sample, no matter whether it is absorption or "apparent absorption", for example, light scattering. As an example, a turbid suspension of light-scattering fat droplet particles such as milk exhibits extinction in the visible part of the spectrum, although it does not necessarily absorb light in the sense of the Lambert–Beer law.

Monochromatic light, one of the prerequisites for the use of the Lambert–Beer law (see above), can be obtained to a decent approximation. The other prerequisites are sometimes difficult to meet. For example, the substance that is analyzed may exhibit fluorescence and consequently re-emit a fraction of the absorbed light at longer wavelengths. Dependent on the geometry of the beam between sample and detector, the intensity I may then be higher than for a non-fluorescing sample, which will lead to an underestimation of the absorption. In this case, dilution of the sample to a concentration where the intensity of the emitted light on the detector is negligible with respect to that of the transmitted light will help.

Biological samples are frequently suspensions, for example, cells, cell organelles, or membranes that scatter light (often even seen by eye). Dependent on the size and form of the suspended particles, a substantial part of the incoming light may be scattered away from the forward direction, which leads to lower intensity at the detector as compared to a clear, non-turbid sample. The intensity reaching the detector varies depending on the active detector area and the distance from the sample and thus on the fraction of the scattered light that reaches the detector. This may vary from photometer to photometer. Since scattering is a wavelength dependent phenomenon, a scattering curve is observed that overlaps with the wavelength dependent absorption. Figure 7.7 illustrates schematically this scattering curve and its consequence for an absorption measurement. In this example we have assumed that the light-scattering particles exhibit dimensions on the order of the wavelength. In practice, these could be membrane fragments, cells, or cell organelles.

A precise analytical description of the scattering curve that would allow subtraction from the measured "extinction" requires information on the properties of the scattering particles and on the medium and cannot be generally given. It depends on the particle size d relative to the wavelength λ and the properties of the medium, for example, its refractive index n. In detail, it also depends on the form of the particle. For spectroscopic measurements in the ultraviolet or the visible spectral range, Rayleigh scattering ($d \ll \lambda$), Rayleigh–Gans–Debye scattering

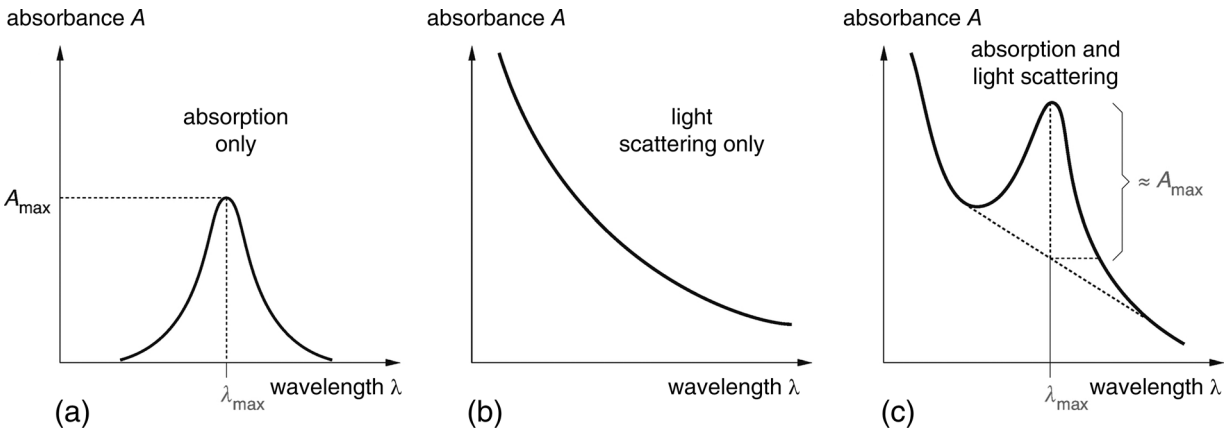


Figure 7.7 Light scattering and its consequence for an absorption measurement. (a) Absorption without light scattering and (b) extinction (apparent absorption due to light scattering particles). In the simplest case they overlap (c) in the measured spectrum. To determine the absorbance they have to be separated by a suitable approximation of the scattering curve.

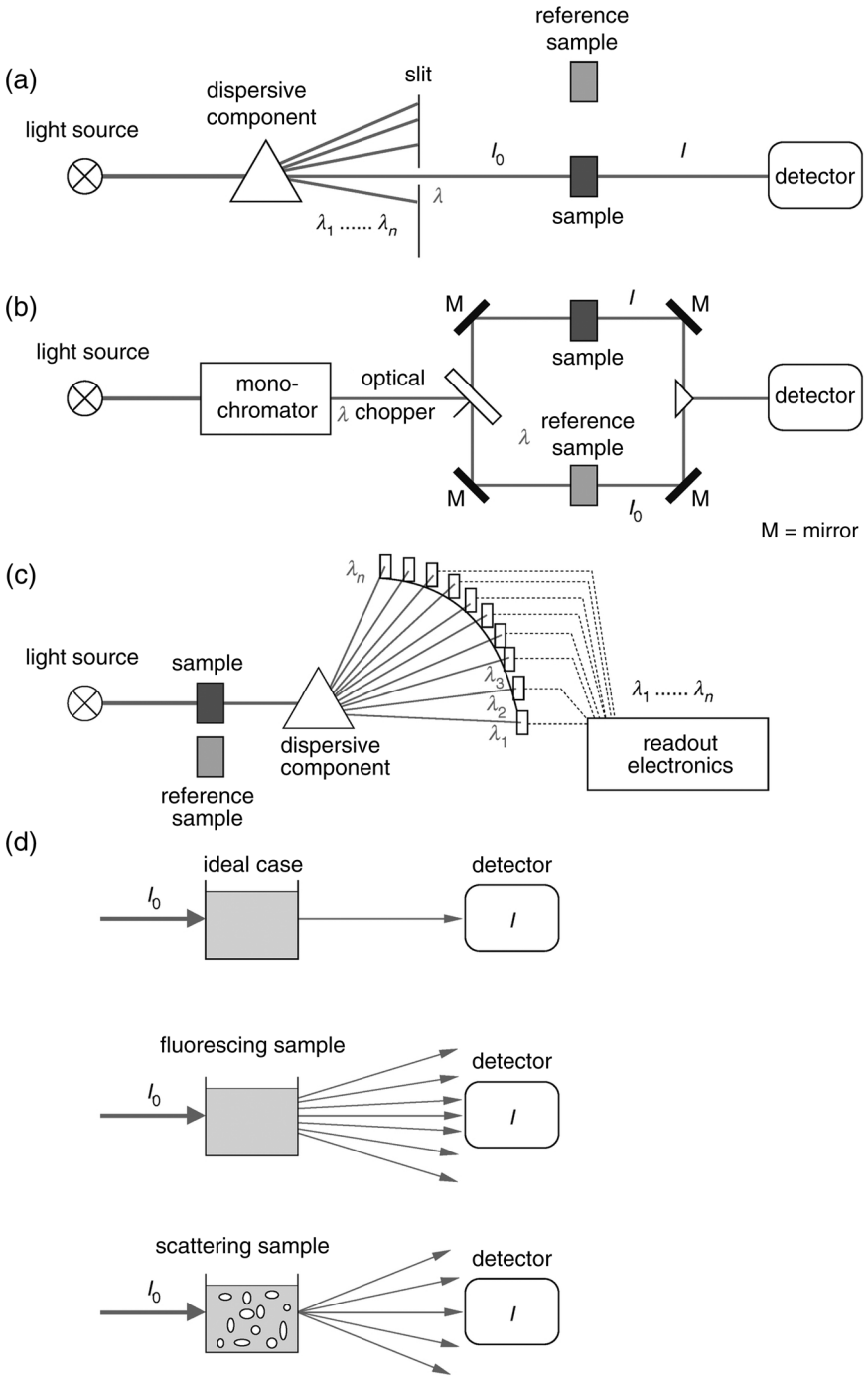


Figure 7.8 Schematic function of photometers and their main sources of error: (a) Single beam photometer; sample and reference have to be recorded separately. (b) Dual-beam photometer; sample and reference are measured alternately. (c) Diode array photometer; sample and reference have to be recorded separately. (d) Main sources of error in photometric measurements (see also shaded text box).

($d < \lambda$), Mie scattering ($d \cong \lambda$), and Fraunhofer scattering ($d > \lambda$) are distinguished. The different nomenclature stands for different theoretical treatments of the scattering process. An example of Rayleigh scattering is the blue sky (scattering of light by molecules of the atmosphere) and of Mie scattering the white clouds in a blue sky (scattering at water droplets of the size of some micrometers).

Correction for Scattering Artefacts For nanoscale particles with a size of up to approximately the wavelength of the light used for analysis, the apparent absorption due to light scattering exhibits a $1/\lambda^4$ dependence, that is, blue light is much more strongly scattered

than red light. The simplest approximation by which to separate absorption from scattering is to use wavelengths outside the absorbance region for the interpolation of the scattering curve and thus for the correction of the absorption value (Figure 7.7c). A judicious arrangement of sample and detector position in the spectrophotometer or the use of special sample chambers can drastically reduce light scattering artefacts.

Frequent errors in photometry (see Figure 7.8d)

- Fluorescent sample, incomplete filling of the cuvette, imprecise adjustment of the cuvette in the spectrophotometer beam;
  Consequences: Detector intensity is too high, absorption apparently too low.
  Remedy: Dilution (for fluorescent samples), masking of the beam that passes sideways through the cuvette.
- Light scattering samples: Intensity is scattered sideways and misses the detector
  Consequences: Detector intensity is too low, absorption apparently too high.
  Remedy: Sample should be mounted close to the detector or a large-area detector should be used; special sample chambers for scattering samples (Integrating spheres, Ulbricht spheres) can be used that collect scattered light even for large scattering angles. The refractive index n of the solvent can be increased to reduce scattering at the particles, for example, by adding non-absorbing macromolecules, such as polysugars, at high concentration.

7.1.4 Photometer

Absorption spectra in the UV, VIS, and NIR range are mostly recorded on routine spectrophotometers. Most of them cover the range 190–900 nm or more. Almost exclusively, these instruments are automatically recording photometers that register transmission or absorbance over a preselected spectral range.

In a dual-beam photometer (Figure 7.8b), the measuring beam is provided by a light source that is a tungsten–iodine lamp for the visible and NIR part (400–900 nm) and a deuterium-filled discharge lamp for the UV part (approx. 200–400 nm). In most cases, switching from one light source to the other by a flipping mirror is done automatically once the wavelength range is set to change. A monochromator, nowadays mostly using a diffraction grating, is used to select the desired wavelength. Because white light is dispersed and separated into its spectral components, the term dispersive spectroscopy is common. An automatic drive on the monochromator is used to scan from an initial wavelength to a final wavelength. The beam is then separated into a sample and a reference beam that pass the respective sample. Both are then recombined at the detector. There are technical variations for this part of the optics. Sometimes a rotating mirror (a "chopper") is used to guide the beam in an alternating way through the sample and the reference branch, detecting them separately; the detector electronics receives a synchronization signal from the chopper to identify I and I_0. An alternative is to split the beam into two approximately equal parts and to detect each one with separate detectors of similar sensitivity. Photomultipliers are frequently used as detectors. They work well for the UV-Vis range but show weak performance for the NIR part; complementarily, silicon photodiodes work up to >1000 nm but exhibit decreased sensitivity for the UV range.

The rapid development of semiconductor light sources, that is, light emitting diodes (LEDs) and lasers, offers the opportunity to build very sensitive, robust, versatile, but nevertheless cost-effective photometers dedicated to specific tasks, such as photometric measurements in clinical settings, spectroscopy *in vivo*, or field measurements in biology. They make use of the high brilliance (emission per area) and stability of semiconductor sources as compared to incandescent lamps and the absence of monochromators with moving parts. Moreover, semiconductor light sources do not exhibit the additional NIR and IR radiation – namely, heat – that incandescent lamps necessarily emit and that could harm a sensitive sample.

A double beam spectrometer that records a sample spectrum and a reference spectrum and calculates from these $A = \log(I_0/I)$ has the advantage that the background absorbance is automatically compensated and the "pure" sample spectrum is obtained directly. This may not be necessary for the UV-VIS part of the spectrum, where water and other solvents are essentially transparent. However, it poses a problem in the NIR and MIR where water and other

solvents exhibit strong absorbance bands; consequently, the spectra of biological materials thus require some sort of background compensation. In addition, the spectral properties of the lamp, the monochromator, the optics, and the detector are also compensated. There is no principal disadvantage of single beam spectrometers that lack the chopper or beam splitter optics. Upon recording a spectrum on a single-beam photometer, the sample spectrum $I(\lambda)$ and reference spectrum $I_0(\lambda)$ are separately recorded, stored, and then used to calculate the absorbance spectrum $A(\lambda)$. The only requirement for this procedure are a highly stable light source and optics.

A common property of both single-beam and double-beam spectrometers is that all spectral elements $\lambda_1 \ldots \lambda_n$ are scanned successively. The advantage of this procedure is that only one spectral element at a time passes the sample and unwanted photoreactions are thus minimized. The disadvantage is the time required for scanning one spectral element after the other, with a minimum time required for each spectral element and the use of only one element at a time, while all other spectral elements are unused. While this is much less a problem for the UV-VIS part of the spectrum where light sources deliver a high photon flux, it is problematic for the MIR part where only a low photon flow density is available from thermal sources.

Multichannel spectrometers can overcome this disadvantage by using many detectors in parallel, each one for a separate spectral element. The working principle of these multichannel devices – which are called diode array spectrometers, optical multichannel analyzers (OMAs), or charge-coupled devices (CCDs), with the name depending on different semiconductor technologies – is shown in Figure 7.8c. All use the simultaneous measurement of many (up to 4096 or even more) spectral elements for the rapid acquisition of spectra. This gain in speed is termed the multiplex advantage. Even with simple, cheap CCD devices (the one-dimensional version of the two-dimension CCD sensor in your digital camera), recording a spectrum may take only some milliseconds, which opens up the possibility to analyze the time course of reactions (e.g., enzyme reactions) from millisecond to seconds. If time resolution is not required, spectra can be averaged to give a much better signal-to-noise ratio.

The optical design of these multiplex spectrometers requires illumination of the sample with white light or at least with light over the entire spectral range selected; the spectral separation of the light is performed after passing the sample. Consequently, the sample is exposed to light in the full spectral range, although only for shorter times. This may result in photochemical reactions or even bleaching of the sample. In addition, light scattering in the sample can lead to a badly defined optics between sample, dispersing element, and multichannel detection.

These optical multiplex analyzers, by design and technology, are single beam photometers; the spectra of sample and reference are thus recorded subsequently. Due to rapid data acquisition, they are often used as detectors for liquid chromatography, where they replace single-wavelength detectors for a better separation of complex mixtures.

The high brilliance (emission per area) of LED and laser emitters that arises from the small emitting area is ideal for coupling into optical fibers to guide light to the sample. The transmitted, scattered, reflected, or emitted light can also be collected with a receiver fiber and guided to a compact CCD detector combined with an integrated grating optics. With this combination, spectrometers for remote measurement in hazardous zones, with immersion probes for the analysis of water quality, or coupled to an endoscope can be implemented.

7.1.5 Time-Resolved Spectroscopy

Spectroscopic methods such as those discussed above are greatly enhanced if they can be performed in a time-resolved manner. For a biochemical reaction thus analyzed – ideally in real time – information can be obtained on the initial and final state of the reaction as well as on the number and identity of reaction intermediates. Reaction rates derived from these time-resolved spectra yield thermodynamic parameters such as reaction enthalpy and entropy.

In principle, any spectrophotometer can be used for time-resolved spectroscopy, simply by successively recording spectra. In this case, the time required to record one spectrum represents the limit for time resolution. It should be significantly shorter than the expected half-time of the reaction. Most commercially available spectrophotometers offer the option to measure the absorption at a fixed individual wavelength (set by the monochromator) as a function of time. Although this procedure does not yield complete spectra, it can be used to monitor the rise and decay of intermediates at different wavelengths. All variants of photometers shown in

Figure 7.8 are suitable for this *single-wavelength analysis*, although the use of a dual-beam photometer causes delays because of the chopper that periodically switches between sample and reference beam. Diode array spectrometers or CCD-spectrometers offer the best time resolution with complete spectra because of the multiplex advantage, and the time frames for the collection of a complete spectrum can be as low as microseconds to milliseconds.

The problem to be solved for rapid time-resolved measurements is usually not the physical time resolution of the spectrometer instrument but rather the procedure for precise reaction control so as to start the process at a given time, the "zero point". In principle, a cuvette suitable for spectroscopic measurements could be used with an enzyme dissolved or suspended in it, and substrate addition could start the reaction. However, even if a homogeneous mixing is performed by stirring, it will take seconds before this reaction mixture can be analyzed by photometry free from turbulence or bubbles.

To start chemical and biochemical reactions by mixing, procedures have been developed that are known as "rapid mixing," "diffusive mixing," or "stopped flow". The techniques use separate thermostated reservoirs for the enzyme and the substrate and syringes to inject these into a reaction chamber equipped with windows for the measuring beam. The volume and form of the reaction chambers are optimized for a rapid and homogeneous mixing. The time limit that can be obtained with these techniques is around 100 µs to 1 ms.

Much faster time-resolved measurements are possible using perturbation techniques. These are based on a sudden alteration of a chemical equilibrium, by a rapid change of pressure or temperature, and the observation of the relaxation processes towards the new equilibrium. Other perturbation techniques use time-resolved spectroscopy with a photochemical triggering of the reaction, for example, by ultrashort laser flashes. The techniques for photochemical excitation and the spectroscopic monitoring of reaction intermediates and products have been developed in recent decades to the picosecond and femtosecond time domain and can thus be used for the direct analysis of intra- and intermolecular dynamics; however, they are limited to photochemically-induced reactions.

Photochemical triggering can indirectly be used to start biochemical reactions. This technique uses specifically designed substrate analogs that have been "inactivated" with a protection group. Although this need not be a steric inactivation, the terms "cage" (for the protection group) and "caged compound" (for the inactivated substrate analog) are used. These "caged compounds" are molecules that can instantaneously release active substrate molecules if the protection group is photochemically activated and split off the substrate.

Time-resolved spectroscopy experiments that involve enzymes and caged substrates to trigger the desired reaction are performed by mixing the enzyme and the inactivated substrate homogeneously, recording spectra of the unperturbed mixture, and then triggering the reaction by exciting the protecting group with an additional light flash. In the case of the frequently used nitrobenzyl protecting group (Figure 7.9), an intense UV flash is used to release a noticeable amount of the substrate. The release step itself involves intermediates and takes nanoseconds to milliseconds, depending on the nature of the protection group. Even if the caged compound is homogeneously distributed among the enzyme molecules and this mixture is homogeneously excited, diffusion of the released substrate to the binding site of the enzyme may take extra time. In ideal cases, the release can even take place directly in the binding site if the caged compound binds despite the steric hindrance due to the protection group.

Figure 7.9 shows the structures of *caged* ATP (a) and *caged* Ca^{2+} (c) that can be used for the rapid release of ATP and Ca^{2+}, respectively. The *caged* neurotransmitter analog (Figure 7.9b) photochemically releases the neurotransmitter carbamoylcholine, and the *caged* proton (Figure 7.9d) can be used to generate a rapid pH jump upon UV excitation. Photochemical excitation of tris(2,2′-bipyridine)ruthenium(II) can induce a rapid photoreduction and was thus called a "caged electron."

To initiate these reactions, the protecting group has to be activated by a short and intense UV flash (in the spectral range 300–350 nm for compounds protected by the nitrophenyl group). The primary photochemistry of these cages is not simple, and not all intermediate reaction steps, side reactions, and by-products have been sufficiently understood. On the other hand, numerous new caged compounds have been synthesized, some of which are even commercially available. It is thus obvious that the rapid mixing and stopped-flow procedures described above will be gradually replaced by photoactivated enzyme reactions and that time-resolved spectroscopic studies on enzymes will be available with much better time resolution.

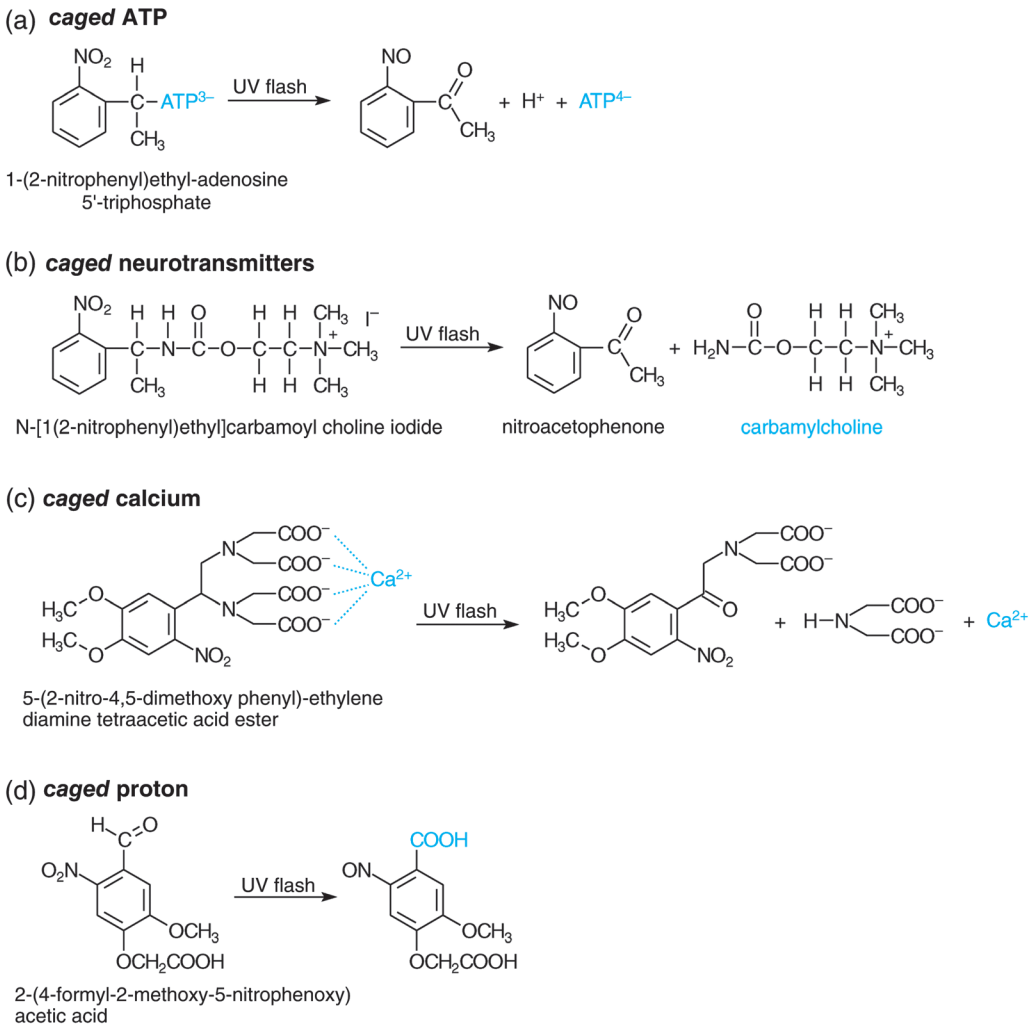


Figure 7.9 Examples of phototriggered effector molecules used to start biochemical reactions. (a) Caged adenosine triphosphate, ATP; (b) caged neurotransmitter analog; (c) caged Ca^{2+}; (d) caged proton used for a pH jump.

Similarly to caged compounds, photoswitches such as azobenzenes can be used to rapidly switch conformations by cis–trans isomerization. In contrast to the caged compounds, which bleach irreversibly, they may be switched back using light at different wavelengths. These bistable switches are increasingly used for the time-resolved spectroscopy of peptides and proteins. They can be inserted as part of artificial amino acids and allow a light-induced switching of peptide conformations, for example, as a trigger of protein folding or unfolding.

7.2 UV/VIS/NIR Spectroscopy

7.2.1 Basic Principles

To apply UV/VIS/NIR spectroscopy for the characterization of proteins we have to distinguish between the absorption of the polypeptide itself, that is, the main chain and the side chains, and the absorbance arising from chromophoric groups, that is, from prosthetic groups, cofactors, bound effector or inhibitor molecules, or pigments.

To illustrate the electronic transitions of the peptide bond and of the side chains of amino acids, we use the scheme shown in Figure 7.4. The basic unit of a polypeptide chain, that is, the

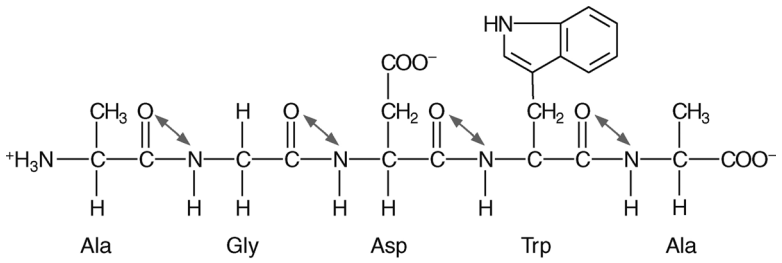


Figure 7.10 Schematic representation of a polypeptide. The electronic transitions of the peptide bond are indicated by arrows.

peptide bond, only contributes in the UV part of the spectrum. In the electronic ground state, the electronic transition dipole of the peptide bond is oriented along an axis connecting the nitrogen atom and the oxygen atom (Figure 7.10). Because of the resonance structure of the peptide bond, a $\pi \rightarrow \pi^*$ transition arises with an absorbance maximum around approx. 190 nm, which can be detected easily because its molar absorption coefficient ε is around $7000\,\mathrm{l\,mol^{-1}\,cm^{-1}}$.

The $n \rightarrow \pi^*$ transition for this bond at around 220 nm is considerably weaker with ε around $100–200\,\mathrm{l\,mol^{-1}\,cm^{-1}}$; in addition, there is an overlap of absorption from the side chains of some amino acids in this spectral region. This transition can thus not be used for an analysis of the polypeptide.

For polypeptides, the $\pi \rightarrow \pi^*$ transitions arising from the side chains of the aromatic amino acids exhibit a pronounced absorption in the range 260–280 nm. The main contributions arise from the amino acid tryptophan ($\lambda_{max} = 280$ nm, $\varepsilon \cong 6000\,\mathrm{l\,mol^{-1}\,cm^{-1}}$), composed of different transitions with maxima between 250 and 300 nm as indicated by the complex structured absorbance spectrum. There are weaker contributions in this spectral range from the $\pi \rightarrow \pi^*$ transition of phenylalanine ($\lambda_{max} = 260$ nm, $\varepsilon \cong 200\,\mathrm{l\,mol^{-1}\,cm^{-1}}$) and from tyrosine ($\lambda_{max} = 275$ nm, $\varepsilon \cong 1500\,\mathrm{l\,mol^{-1}\,cm^{-1}}$).

These transitions arising from the aromatic amino acid side chains in a polypeptide can be used for a simple yet not very precise measurement of protein concentration. If one assumes an average distribution of the amino acids tryptophan and tyrosine in a polypeptide (neglecting the contribution from phenylalanine), a polypeptide solution at a concentration of $1\,\mathrm{mg\,ml^{-1}}$, measured at a path length of 1 cm, exhibits an absorption $A_{280} \cong 1$. If this quick test is used without any refinement, that is, for many very different polypeptides and proteins, the error is frustrating: a $1\,\mathrm{mg\,ml^{-1}}$ protein solution at 1 cm path length exhibits $A_{280} \cong 1 \pm 0.5$. Although this may appear a very low precision, notably, this method is non-destructive, which means that the protein solution can be used for further processing or analysis, and other methods, such as staining by the Lowry technique, are not much more precise. The precision can be significantly increased if the number of tyrosines and tryptophans for the protein under investigation are precisely known. Another very popular use of this quick spectroscopic analysis is the check for purity for proteins with a chromophore group by measuring the ratio of the chromophore absorption in the visible part of the spectrum and the UV absorption at 280 nm. Any contamination with other proteins would show in an increased 280 nm absorption.

Lowry Assay, Section 2.1.2

7.2.2 Chromoproteins

Based on the above section on the intrinsic electronic absorption of polypeptides, all proteins without a chromophoric group, a pigment, or a cofactor absorbing visible light will exhibit their lowest-energy electronic transition in the UV, that is, they appear colorless to the human eye. The apparent yellow color of albumin, the dominant protein in blood plasma, is simply caused by the UV absorption of its aromatic residues tailing out to the blue-violet part of the visible light spectrum.

An exception to this rule is formed by a small group of polypeptides that lack any colored cofactor but nevertheless exhibit absorption in the near-UV or visible part of the spectrum. Chromophoric groups may form after protein folding by a cyclic reaction of several side chains to yield an intrinsic chromophore, without the involvement of additional cofactors. This group of polypeptides is rather small but includes green fluorescent protein and its derivatives that will be dealt with below.

Many proteins carry one or more cofactors with electronic transitions in the visible part of the spectrum and thus appear colored – these proteins are thus called chromoproteins. These visible-light absorbing cofactors are sensitive probes for the function of the protein, because they are often directly involved in the catalytic function of the protein or enzyme and change their electronic properties concomitant with the catalytic reaction. These chromoproteins include as cofactors either atoms or smaller molecules; they range from metal centers via flavins, heme molecules, to chlorophylls or retinals. Their spectroscopic properties cannot be described here comprehensively; we thus give some specific examples.

Rhodopsins A well-known group of chromoproteins is the rhodopsins, which contain as a chromophoric group retinal, a vitamin A aldehyde, in different isomeric forms. The most prominent representative is the visual pigment rhodopsin. There are also many different bacterial analogs that either act as a proton pump, such as bacteriorhodopsin, which enable light orientation in halobacteria, or enable ion channels to open or close upon illumination, such as channel rhodopsin.

In these proteins, retinal is bound covalently via a protonated Schiff base to the ε-amino group of a lysine residue. Apart from this covalent binding, retinal is subjected to further interactions with amino acid side chains lining its binding pocket. In all retinal proteins, the binding and local interactions cause a substantial redshift (bathochromic shift, see Section 7.1.2) of the absorption maximum. Free 11-*cis* retinal in ethanol absorbs at around 360 nm, while the same retinal in the visual pigment rhodopsin absorbs at 500 nm in the dark state. Only a fraction of this shift is due to the covalent binding of retinal to opsin (the retinal-free apoprotein) via the Schiff base. The local interactions with the polyene chain and with the β-ionone ring in the binding pocket contribute significantly to the redshift.

The strong electronic transitions of retinal in rhodopsin that lead to intense bands are caused by delocalized π electrons and can be used as sensitive spectroscopic probes for the actual state of retinal in its binding site and thus for the actual state of the protein. Figure 7.11 shows the structure of retinal in its 11-cis form and its binding to the ε-amino group of a lysine residue together with the absorption spectrum in the free form in solvent and the opsin-bound form.

If rhodopsin is illuminated by light in the spectral range around 500 nm, absorption of light leads to a photochemical reaction with subsequent thermally activated dark reactions. This leads to different intermediates of rhodopsin that are summarized in a simplified scheme in

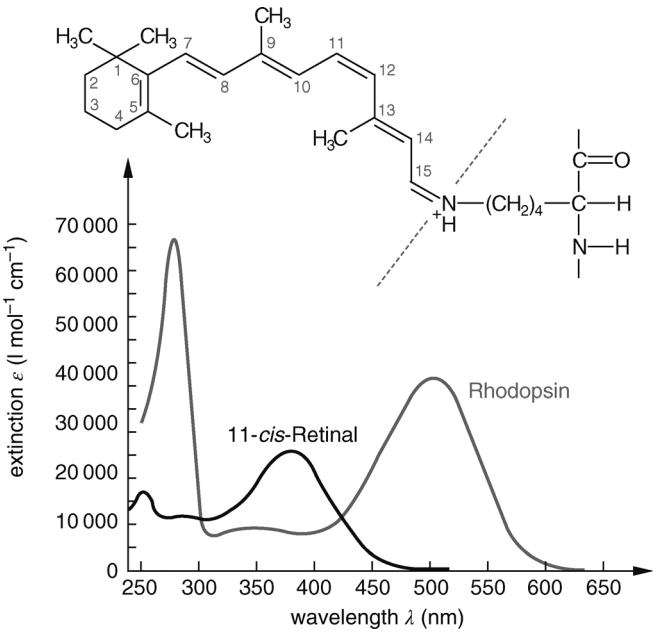


Figure 7.11 Structure of retinal in the 11-cis form bound to the ε-amino group of a lysine residue via a protonated Schiff base. This binding of retinal in the visual pigment rhodopsin leads to a strong redshift in the absorption spectrum.

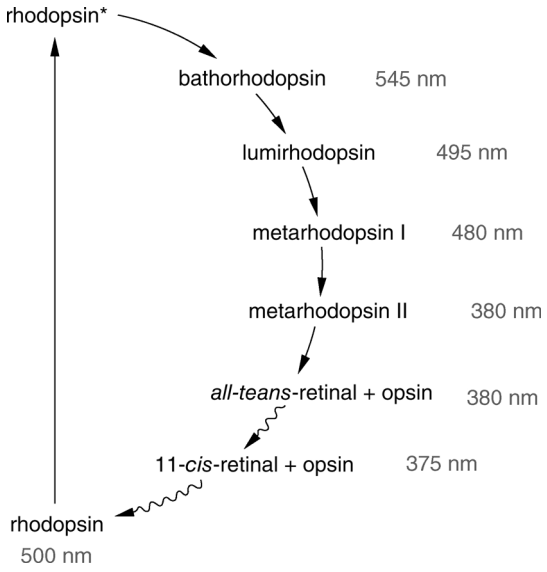


Figure 7.12 Simplified reaction sequence of rhodopsin after absorption of a photon. The wavelength indicated for each intermediate indicates the respective absorption maximum. Wiggly lines are for reactions that do not occur spontaneously but require enzymatic activity.

Figure 7.12. The reaction step relevant for the coupling of the photoreaction and the dark reactions to the processes of biochemical amplification and regulation, and thus to what is termed as "vision", is the equilibrium reaction from metarhodopsin I (MI) to metarhodopsin II (MII). All intermediates within this reaction sequence, frequently called "bleaching", differ by their absorption maximum and absorption strength and can thus be characterized by spectroscopic measurements. One possibility is to follow the rise and decay of intermediates by time-resolved spectroscopy, which requires picosecond time resolution for the first intermediates and even micro- and millisecond time resolution for the later reaction steps. As an alternative, spectra of intermediates can be obtained by illumination at low temperature (e.g., 77 K) followed by gradual warming to see the thermally activated reaction intermediates evolve.

The first, redshifted, intermediate, termed *bathorhodopsin* ($\lambda_{max} = 545$ nm), exhibits stronger delocalization of the π electrons in the polyene chain of retinal. The strong blue-shift of the chromophore absorption in the intermediate *metarhodopsin II* ($\lambda_{max} = 380$ nm) can be explained by deprotonation of the Schiff base and a structural opening of the protein. In a final step of the reaction sequence, retinal is detached from opsin and the absorption of free retinal is observed.

Analogous reaction sequences are observed for other retinal proteins. Bacterial rhodopsins undergo light-driven cyclic reactions. An example is bacteriorhodopsin, a light-driven proton pump from the archaebacterium *Halobacterium salinarium*. In bacteriorhodopsin, retinal in the all-trans form is bound to the side chain of a lysine residue at position 216. Upon absorption of a photon, a cyclic reaction is started that overall takes a few milliseconds and involves conformational changes in the protein scaffold as well as internal proton transfer processes. This is all driven by the light energy absorbed by retinal and results in a net transfer of a proton from the inside to the outside of the cell.

Cytochromes Further examples of chromophoric groups of proteins that lead to characteristic absorption in the visible part of the spectrum are porphyrins and hemes. Within the huge class of heme proteins, the water soluble cytochromes *c* and c_2 are among the smallest representatives; they are ubiquitous as electron shuttles in many electron transfer chains, for example, in the respiratory chain or in photosynthesis. In the case of cytochrome *c,* the prosthetic group is a heme molecule covalently bound to cysteine side chains via thioether groups in a -Cys-X-Y-Cys- heme binding amino acid sequence within the polypeptide chain (Figure 7.13). As well as these covalent bonds, there are additional interactions between heme and protein via the axial ligand of the iron and the propionate residues of the heme, as well as by polar or nonpolar residues in the heme binding pocket. These specific interactions lead to substantial differences in the spectral and redox properties of these heme proteins, albeit for

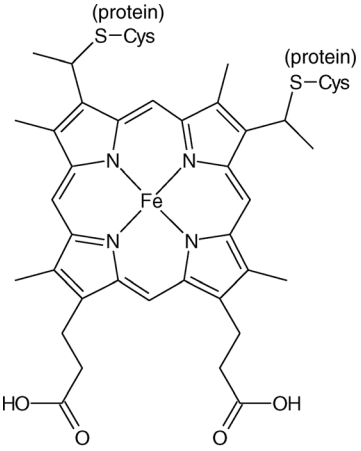


Figure 7.13 Structure of the heme group of cytochrome c bound to cysteine side chains.

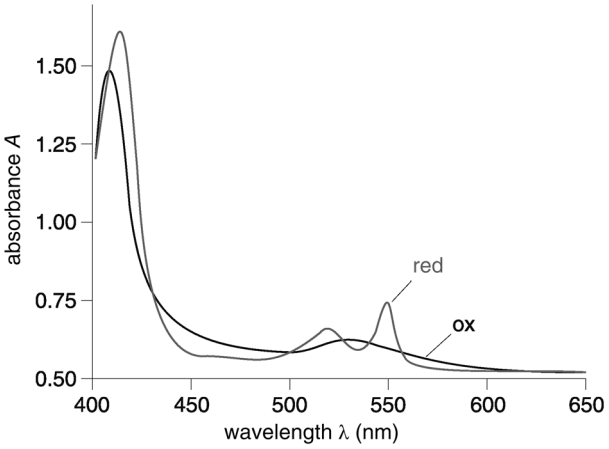


Figure 7.14 Absorption spectra of oxidized and reduced cytochrome *c* from horse heart and the difference spectra calculated for the full transition from the oxidized to the reduced state (reduced-minus-oxidized) and vice versa. The redox transitions were obtained at a transparent electrode in a spectroelectrochemical cell.

identical heme molecules, and facilitate the function of different *c*-type cytochromes in different parts of biological electron-transfer chains.

The optical absorption spectrum of cytochrome *c* is characterized by an absorption band at around 550 nm (the so-called α-band), a broader and somewhat weaker band at approx. 530 nm, and a strong band at approx. 400 nm, the Soret band (Figure 7.14). All three bands depend on the redox state and can thus be used for its assessment, in the isolated protein as well as in entire membranes, organelles, or cells.

Figure 7.15 illustrates the redox reaction of cytochrome *c* in a spectroelectrochemical cell. A transparent electrode is used to apply a certain potential; the measuring beam monitors the absorbance of the proteins on or near the electrode, and counter and reference electrodes are used to adjust the correct potential. An absorbance spectrum recorded at a given redox potential reflects the fractions of reduced and oxidized cytochrome *c* molecules. For a better presentation of the redox-induced alterations, the difference spectra with respect to a fully oxidized (or reduced) protein are calculated and displayed, as seen in Figure 7.15b.

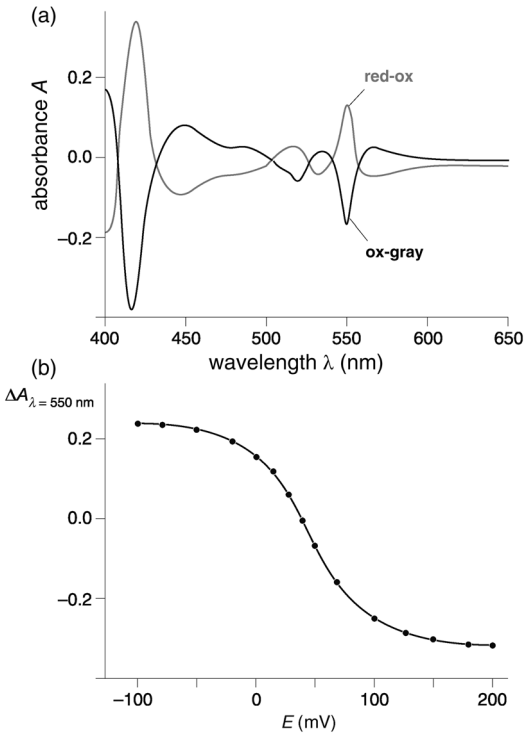


Figure 7.15 (a,b) Redox titration of cytochrome *c* measured at the α band at 552 nm. The signals are taken from difference spectra as shown in Figure 7.14, but for different potential steps. The full line in (b) represents a Nernst function for $n = 1$.

If the amplitude of the difference signal at a given wavelength is plotted against the applied potential, a redox titration curve is obtained that can be described by a Nernst function:

$$E = E_0 + \frac{RT}{nF} \cdot \ln\left(\frac{c_{ox}}{c_{red}}\right) \tag{7.9}$$

where: E is the applied or (in equilibrium) measured potential;
E_0 is the midpoint potential with equal concentrations for the oxidized (ox) and reduced (red) species;
R the gas constant ($8.314\,\mathrm{J\,K^{-1}\,mol^{-1}}$);
T is the absolute temperature in K;
F is the Faraday constant ($96485\,\mathrm{C\,mol^{-1}}$);
n the number of electron transferred.

By fitting this Nernst function to the measured absorbance values, the midpoint potential E_0 and the number of transferred electrons n can be determined (in the case of cytochrome c, $n = 1$).

Redox titrations like this one can also be performed chemically, by adding reductants or oxidants, and by assessing the redox state of cytochrome c from the absorption spectra. However, the advantage of an electrochemically performed redox titration is the much more precise adjustment of potentials and the fact that no dilution by oxidants or reductants occurs. Midpoint potentials thus determined are much more precise and allow us to, for example, measure the influence of polarity or charge of individual amino acid residues on the redox properties of the heme, by comparing native cytochrome c with site-directed mutants where these residues are altered.

Metalloproteins A chromophoric group can also be formed by one or more metal ions bound or complexed in the protein. Some examples include non-heme iron in photosynthetic reaction centers, manganese in the water splitting complex from cyanobacteria or plants, and copper in heme-copper oxidases or in small water-soluble redox proteins such as azurin or plastocyanin. These metal ions may have magnetic properties because of their nuclear or electron spin, and thus may be interesting for bioanalytical studies using EPR or NMR methods. Many of these metal centers also exhibit distinct electronic levels due to binding in the protein, with transitions in the UV, visible, or NIR part of the spectrum.

The deep blue color of the water-soluble blue copper proteins (one of them is azurin: the color gave the name) in the oxidized state is an example of these electronic transitions in metalloproteins. The electronic transition of these blue copper proteins with an absorbance maximum around 600 nm and a moderately strong molar absorption coefficient ($2000–5000\,\mathrm{l\,mol^{-1}\,cm^{-1}}$) arises from a charge transfer between the ligands and the copper; in azurin the strongest of these so-called charge-transfer bands arises from the sulfur of the cysteine ligand.

Figure 7.16 shows schematically the coordination of the copper ion by two histidine and two cysteine side chains for azurin. A peptide carbonyl acts as a fifth ligand. Figure 7.17 shows the

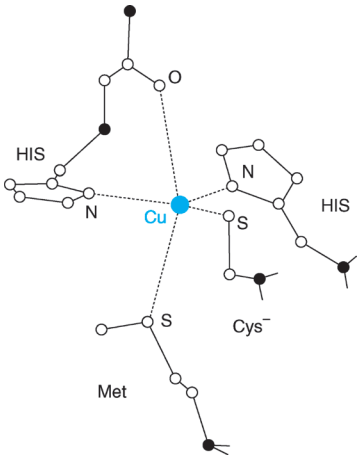


Figure 7.16 Coordination of the copper ion by amino acid side chains and a peptide carbonyl for the blue copper protein azurin.

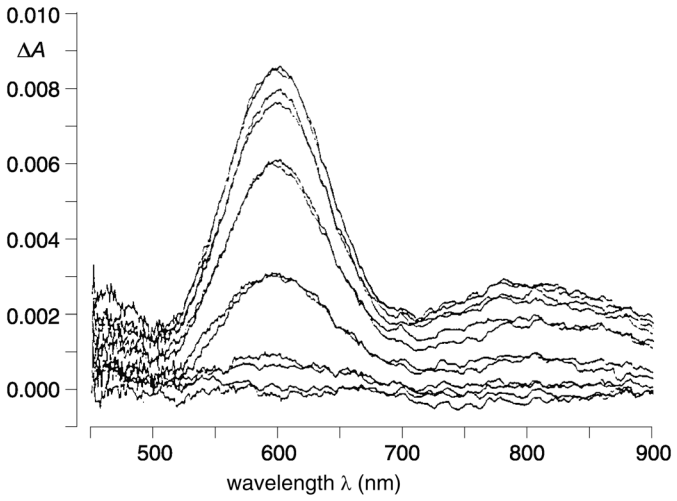


Figure 7.17 Absorption spectra of the archaebacterial blue copper protein halocyanin at different redox potentials.

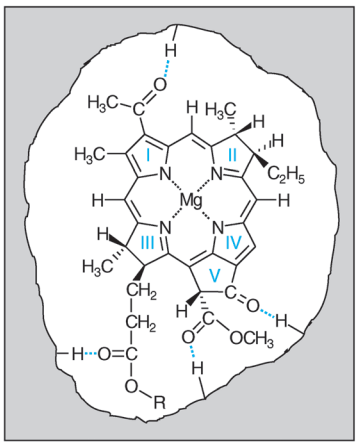


Figure 7.18 Structure of a hypothetical fully bonded bacteriochlorophyll *a* molecule in a protein binding pocket. The carbonyl groups at the periphery of the molecules can form hydrogen bonds with amino acid side chains in the protein (dotted lines). The two axial ligands of the magnesium are not shown.

spectra of another representative of these blue copper proteins, namely, halocyanin from a halophilic archaebacterium, at different redox potentials. In the reduced state, these proteins are colorless. As the potential increases, the charge-transfer band appears.

Alterations at the geometry of the metal–ligand center or of the hydrogen bonding properties can change the local electron density at the copper atom; as a consequence, the electronic levels and the redox potentials may change. In any of these cases, the electronic transition can be used to characterize the redox state.

Chlorophylls Chlorophylls and chlorophyll–protein complexes are among the strongest absorbing biological molecules and exhibit very intense electronic transitions. Actually, around 60 natural variants of chlorophyll molecules are known and have been characterized by their structure and electronic properties. The common feature of chlorophylls is the cyclic tetrapyrrole structure and the central magnesium coordinated by the four tetrapyrrole nitrogens. The chlorophyll variants differ in the peripheral substituents, which influence all or part of the conjugated system. Thus, these groups influence the position and intensity of the electronic transitions as well as the redox properties of chlorophylls.

In biological structures, chlorophylls are non-covalently bound and are always associated with a protein (the possible exception to this rule is the chlorosomes of green photosynthetic bacteria). This non-covalent binding can be nicely demonstrated by extracting chlorophylls (and other pigments) from leaves with an acetone–water mixture. Most protein complexes that bind chlorophylls are transmembrane proteins containing histidine as the fifth or sixth (axial) ligand of the central magnesium. The binding pockets for chlorophylls in these proteins consist of a nonpolar region in the range where the tetrapyrrole ring is located and more polar amino acid residues for the region where the peripheral chlorophyll groups such as the R—C═O bonds are placed. These acetyl, keto, or ester C═O groups can efficiently form H-bonds with partners in the protein.

Figure 7.18 illustrates chlorophyll binding and interaction for a hypothetical "fully bonded" bacteriochlorophyll *a* from photosynthetic bacteria. Except for the fifth and sixth axial ligand of the magnesium, which would lie above and below the plane of the figure and are thus not shown, all carbonyls (2a-acetyl at ring I, 9-keto at ring V, 10a ester and 7c ester at ring V and at the connection to the phytol chain) are hydrogen-bonded to partners in the protein.

Figure 7.19 shows the absorbance spectra of free bacteriochlorophyll *a* and *b* in ethanol (Figure 7.19a) as well as spectra of protein-bound bacteriochlorophyll *a* in different light-harvesting pigment–protein complexes from photosynthetic bacteria (Figure 7.19b). The lowest-energy singlet transition (Q_y, $S_0 \rightarrow S_1$) at the red end of the spectrum ($\lambda_{max} \cong 770$ nm) is very strong for both bacteriochlorophyll *a* and *b,* with a molar absorption coefficient ε of $100000\,l\,mol^{-1}\,cm^{-1}$. Transitions with higher energy are located around 590 nm (Q_x, $S_0 \rightarrow S_2$) and at 380–400 nm. The orientation of the transition dipole moments is in the plane of the tetrapyrrole ring along an axis from ring II to ring III (termed Q_x) and an axis from ring I to rings IV/V (termed Q_y). Chlorophylls are essentially planar molecules, and any slight doming (i.e., moving the magnesium outside the plane) distorts the electronic levels. The orientation of the transition moments Q_x and Q_y in the molecule geometry allows one to "fine tune" electronic levels and redox levels by providing suitable interactions at the chlorophyll periphery. It also explains why the Q_x transition can only be tuned in a minor way because H-bonding is not possible (conjugation does not extend beyond ring III), while the Q_y transition can be efficiently tuned because of multiple H-bonding effects to the conjugated system that extends from ring I to rings IV and V.

The basic features of the chlorophyll absorption spectrum with three energetically separate transitions are conserved approximately upon binding of bacteriochlorophyll in the protein matrix. The most drastic changes occur for the $S_0 \rightarrow S_1$ transition. Its energy is considerably lowered and gives rise to absorption bands at approx. 800, 850, or 875 nm for different light-harvesting complexes. For the structurally related bacteriochlorophyll *b,* the lowering of the lowest singlet transition is even more dramatic: while the free pigment in ethanol absorbs around 790 nm the same transition is strongly shifted to the near-IR at 1020 nm in the protein complex.

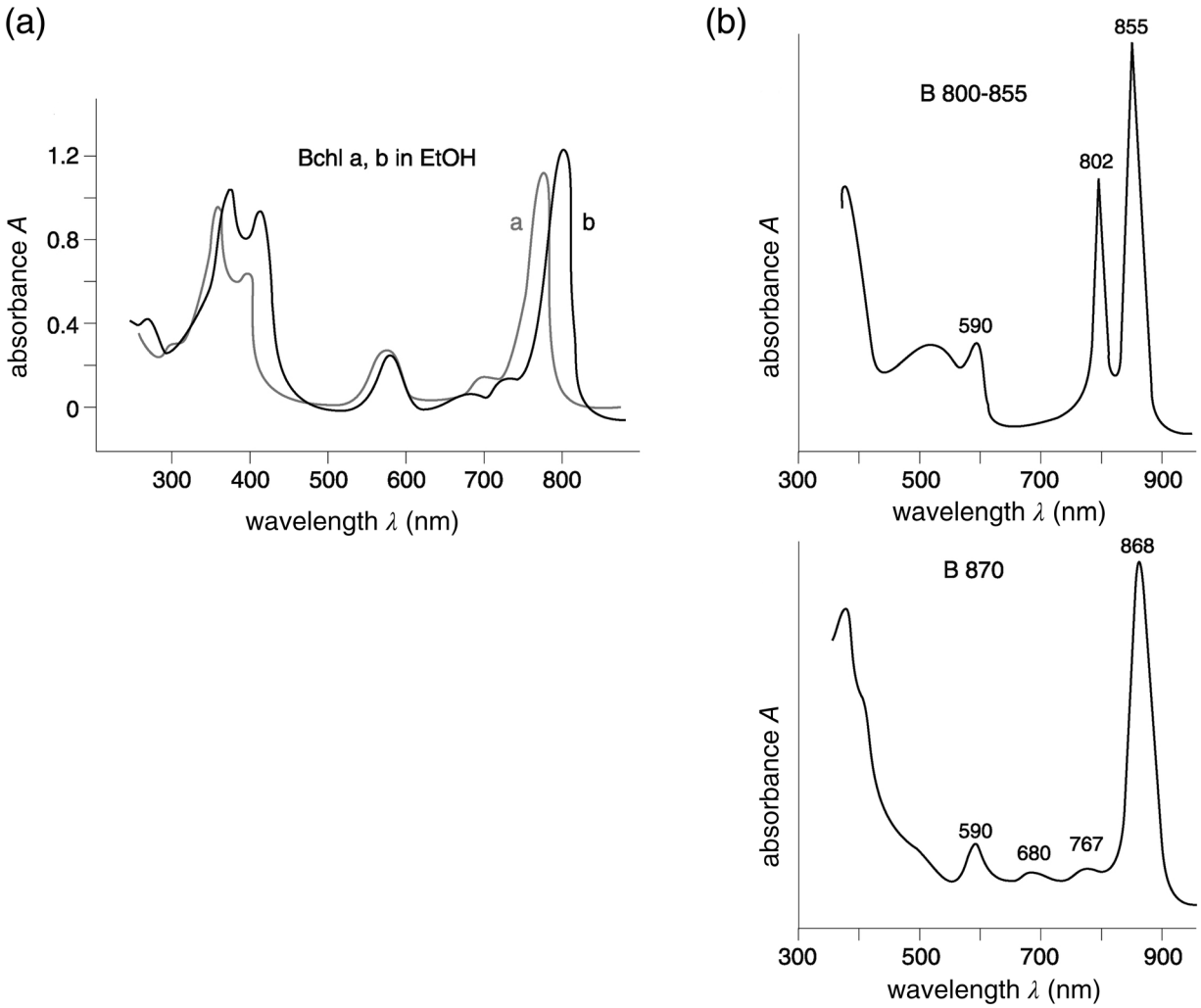


Figure 7.19 (a) Absorption spectra of bacteriochlorophyll *a* and *b* in ethanol (BChl = bacteriochlorophyll); (b) absorption spectra of bacteriochlorophyll *a* in two different pigment–protein light-harvesting complexes (B 800-850 and B 875) from photosynthetic bacteria.

Despite great success in the crystallization and structural analysis of chlorophyll–protein complexes in photosynthesis, the details of all interactions that determine the precise position and intensity of the electronic transition are far from being fully understood, and cannot be derived completely from the electronic spectra. However, these spectra can be used to characterize the type and quantity of the pigment–protein complexes and to study their reactions. This is facilitated by the strong transitions. The absorption spectra furthermore offer sensitive tests for the intactness of the pigments and proteins in the course of isolation and purification. Even minor degradations of the protein or oxidation processes of the pigments lead to strong blue-shifts with respect to the native and intact complex.

In many cases, electronic transitions in biomolecules can be found in the UV, VIS, or NIR range that can be used for quantification *in vitro* or *in vivo*. Almost always, the absorption maximum and/or the intensity of absorption energy of the transition offer the possibility for further analysis of structural, functional, and, with time-resolved techniques, dynamic properties. Finally in this section dealing with the electronic transitions of biological macromolecules structures, the major transitions of natural chromophores are summarized in Table 7.2.

Table 7.2 Electronic absorption properties of chromophores in biological macromolecules.

Range	λ_{max} (nm)	ε (l·mol^{-1}·cm^{-1})	Transition type, absorbing group
UV	≈190	≈7000	$\pi \rightarrow \pi^*$ Transition of peptide bond
	≈220	≈100	$n \rightarrow \pi^*$ Transition of peptide bond
	260	≈13000	$n \rightarrow \pi^*$, $\pi \rightarrow \pi^*$ Adenine
	275	≈6000	$n \rightarrow \pi^*$, $\pi \rightarrow \pi^*$ Guanine
	267	≈6000	$n \rightarrow \pi^*$, $\pi \rightarrow \pi^*$ Cytosine
	264	≈8000	$n \rightarrow \pi^*$, $\pi \rightarrow \pi^*$ Thymine
	258	≈6600	$n \rightarrow \pi^*$, $\pi \rightarrow \pi^*$DNA
	257	≈200	$\pi \rightarrow \pi^*$ Aromatic side chain, phenylalanine
	274	≈1400	$\pi \rightarrow \pi^*$ Aromatic side chain, tyrosine
	280	≈5600	$\pi \rightarrow \pi^*$ Aromatic side chain, tryptophan
VIS	≈420	≈125000	Soret band of hemes
	≈450	≈120000	β-Carotin
	500	≈42000	Retinal in the visual pigment rhodopsin
	550	≈18000	α Band of hemes in cytochrome, reduced-minus-oxidized
	590	≈25000	Q_x transition of bacteriochlorophyll (in ethanol)
	≈600	≈5000	$\pi \rightarrow \pi^*$ Transition in the flavin radical
	590–630	≈2000–5000	Charge-transfer band in type I copper proteins
	772	≈100000	Q_y Transition of bacteriochlorophyll *a* (in ethanol)
NIR	800	>100000	Q_y Transition of bacteriochlorophyll *a* in antenna protein complexes
	850	>100000	Q_y Transition of bacteriochlorophyll *a* in antenna protein complexes
	860	≈80000	Bacteriochlorophyll dimer in photosynthetic reaction centers
	1020	>100000	Q_y Transition of bacteriochlorophyll *b* in antenna protein complexes

7.3 Fluorescence Spectroscopy

7.3.1 Basic Principles of Fluorescence Spectroscopy

The term scheme (Figure 7.6) allows different ways for the relaxation of an electronically excited molecule to the electronic ground state. Apart from radiationless decay, where the energy difference is dissipated in vibrational energy of higher excited vibrational states and finally ends up in heat, there is the possibility of electron transfer and a variety of photochemically-initiated chemical reactions. This higher reactivity of molecules in the photochemically excited state of a molecule can be easily explained by the presence of an electron in a higher orbital with a different – typically lower – stability.

The most relevant process of deactivation for bioanalytical spectroscopy, however, is the emission of a photon. To discuss these phenomena we will use the term scheme shown in Figure 7.6. Light emission can take place in the singlet system as fluorescence and from the triplet system as phosphorescence. Of these two, and of a third one called delayed fluorescence, the measurement of fluorescence is a routine technique applied for many bioanalytical studies.

As discussed in Section 7.1, electronic excitation of a molecule starts from the lowest energy singlet state S_0 to the first, second, or higher excited singlet state, and there, according to the Franck–Condon principle, in the vibrational state v_n with best matching of the nuclear positions. If the molecule is excited to the second or higher singlet state (S_2 or S_n), radiationless relaxation leads rapidly, in around 10^{-13} s, to the lowest vibrational state $v = 0$ of the first excited singlet state, S_1. Direct emission of a photon from a higher excited singlet state is a rare exception.

Relaxation from the state S_1, $v = 0$ can take place radiationlessly via internal conversion. An explanation for this deactivation process is simply that higher excited vibrational states v_n of the

electronic ground state S_0 can reach energies as high or higher as the state S_1. Decay via emission of a photon is termed fluorescence (Figure 7.6).

In Figure 7.6, the length of the arrow representing fluorescence is considerably shorter than that for absorption. Since this length is a measure for the energy of the absorbed and emitted photon, this corresponds to a redshift of fluorescence in the spectrum of the molecule with respect to its absorbance. For an explanation, we consider the occupation and the relaxations of the vibrational levels for absorption and emission. Absorption from the ground state S_0 takes place from an occupation of the vibrational levels that is given by temperature. For room temperature, this is almost exclusively from $v = 0$. According to the Franck–Condon principle (Section 7.1), the absorption process leads to the same ($v = 0$) or a higher ($v = 1, 2, \ldots$) excited vibrational level in S_1. A higher vibrational excitation ($v = 1, 2, \ldots$) principally corresponds to the molecule being at a higher temperature, and thus requires cooling to the thermal equilibrium. It is common to use the term "thermalization" or "vibrational cooling" for this phenomenon. As a consequence of these reactions given by the term scheme, fluorescence exhibits the following properties:

- the fluorescence spectrum of a molecule is redshifted with respect to the absorption spectrum;
- the emission spectrum is independent of the absorption spectrum;
- the structure of the vibrational levels determines the structure of the absorption and the emission spectrum: the vibrational substructure of S_1 determines the absorption spectrum, while the vibrational structure of S_0 determines that of the emission spectrum.

Relaxation from S_1 to S_0 occurs via emission of radiation from a state that has a relatively long lifetime, of the order of 10^{-9}–10^{-7} s. This renders fluorescence studies highly attractive for biological molecules, because an abundance of biologically relevant processes occur within this time regime. However, this lifetime is only valid in the absence of alternative decay pathways, that is, if radiationless deactivation, intersystem crossing, as well as electron transfer and chemical reactions starting from the S_1 state can be neglected.

Typical fluorescence lifetimes are of the same order of magnitude, or even longer, as the time for diffusion, rotation, or conformational changes of biological macromolecules. Consequently, a study of these processes by fluorescence spectroscopy is promising. As a first step, it should be clarified whether fluorescence measurements can be performed directly on the molecule, that is, with its own ("intrinsic") chromophores, or whether it will be necessary to introduce additional ("extrinsic") chromophores as fluorescent probes, either into the molecule or the structure (e.g., a lipid membrane) to be studied.

Fluorescence spectroscopy of proteins with intrinsic probes is limited to the fluorescence of the amino acid tryptophan. As outlined in Section 7.2, the side chains of aromatic amino acids exhibit absorption bands in the UV, with that of tryptophan as the strongest one at around 280 nm (see also Table 7.2). Fluorescence from this electronic transition is observed at around 300–350 nm. In a polar environment, this fluorescence is redshifted by some nanometers. Measurement of the emission spectrum of tryptophan can thus be used to draw conclusion about the local environment of the emitter. For example, a redshift of tryptophan fluorescence is observed in the course of protein unfolding because well-shielded side chains gradually become exposed to water. Another application of this intrinsic fluorescence is the digestion of proteins by proteases, where typically a redshift of tryptophan fluorescence is observed as the residues become more accessible, and subsequently a blue-shift can be detected if proteolysis products form aggregates because of hydrophobic interactions.

For molecules in polar solvents, a further redshift of fluorescence called solvent relaxation is observed in addition to the redshift discussed above that is caused by vibrational relaxation. During the time of photon absorption (10^{-15}–10^{-18} s) solvent molecules are oriented according to the polarity or charges of the molecule. During the much longer lifetime of the excited singlet state, solvent molecules can reorient and form an energetically favorable solvent–molecule structure, thus resulting in a further redshift. A third reason for a redshift, due to experimental inadequacies, will be discussed below.

The intensity of fluorescence depends on the concentration of the molecule investigated:

$$F = I_0 \phi (2.303 \varepsilon c d) \tag{7.10}$$

where F is the fluorescence intensity, I_0 is the intensity of the exciting light, and ϕ is the fluorescence quantum efficiency, that is, the ratio of emitted to absorbed photons. The parameters ε, c, and d define the absorption according to the Lambert–Beer law (Equation 7.7).

For an increasing concentration of the emitting absorbing/molecule, changes of intensity and a redshift of emission are observed. The reason for these effects is termed quenching of fluorescence. Its physical reason is that energy from an excited molecule is directly transferred to another molecule. This reduces the probability of emission and allows radiationless decay of the excitation. This process is concentration dependent and relies on the formation of molecular aggregates or simply the collision of molecules. If a fluorescence experiment is designed to determine quantum efficiencies, this concentration dependency has to be considered very carefully.

7.3.2 Fluorescence: Emission and Action Spectra

Figure 7.20 shows the two possibilities by which to perform a fluorescence experiment. For routine applications, it is sufficient if the fluorescent molecule is excited with light at an appropriate wavelength, λ_{exc}, and the intensity of the emitted light is measured at a second wavelength, λ_{em}. It is normally sufficient to limit the excitation wavelength range with a filter to the effective part of the absorption spectrum of the sample. To prevent scattered excitation light from reaching the detector, a second filter is used in front of the detector. Ideally, both filters are complementary to block the excitation light. Measuring the emitted light at 90° with respect to the incoming excitation beam is a common setup and further helps to prevent excitation light from reaching the detector; however, this is not mandatory.

A more thorough analysis of fluorescence may require information on the emission maximum or on the structure of the emission band. In this case, the filter in front of the detector has to be replaced by a monochromator. Ideally, emission spectra can be recorded in this configuration at wavelengths near the fixed excitation wavelength. Upon recording an emission spectrum, care should be taken to consider the spectral transmission of the optical components and of the monochromator, and especially the wavelength-dependent sensitivity of the detector. Any measurement of the emission intensity versus the wavelength, in the first instance, will only yield relative fluorescence intensities as a function of the measuring wavelengths. These intensities require correction with the help of fluorescence standards, that is, with samples of known quantum efficiency and emission spectrum. This "quantum corrected emission spectrum" can then be evaluated quantitatively.

In some cases it may be necessary to characterize the transfer of excitation energy within a set of chromophores. This can be done by recording an action spectrum: The intensity of the emitted

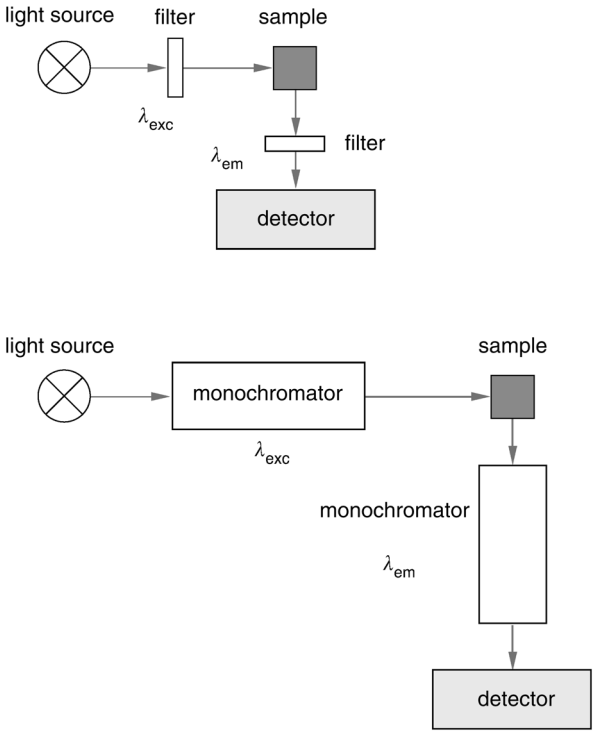


Figure 7.20 Schematic setup for fluorescence studies. In the simplest case, a sample is excited with a fixed wavelength (λ_{exc}) and the fluorescence intensity is measured at a fixed emission wavelength (λ_{em}). These wavelengths are selected by complementary filters. If these filters are replaced by a tunable filter or a monochromator, emission spectra can be recorded by setting λ_{exc} at a fixed value and recording the emission intensity as a function of the emission wavelength λ_{em}. Action spectra are recorded with fixed emission wavelength λ_{em} by scanning the excitation wavelength λ_{exc} and recording the emission intensity.

light is recorded at a fixed emission wavelength λ_{em}, and the excitation wavelength λ_{exc} is varied. This use of fluorescence is known as *fluorescence action spectroscopy*; for a single chromophore ("fluorophore") the action spectrum should correspond to the absorption spectrum.

Light sources for fluorescence measurements should deliver high light intensities in the short wavelength range of the spectrum. Typically, high-pressure xenon lamps are used; tungsten iodine lamps are rather uncommon, unless excitation in the orange–red part of the spectrum is required. Laser excitation is quite common, in particular if time-resolved fluorescence measurements, for example, for fluorescence lifetime measurements, are planned. With the advent of intense and high-power light emitting diodes (LEDs) in the blue, violet, and UV part of the spectrum and semiconductor lasers, cheap and compact excitation sources for fluorescence measurements have become available. By modulating or pulsing the electric current driving these semiconductor LEDs or lasers, time-resolved measurements are easily performed.

For the detection of fluorescence, photomultiplier tubes can be used with high sensitivity in the UV or in the blue–green region of the VIS spectrum. Their amplification can be driven high enough to yield a current output even for the detection of single photons; this mode is called the *photon counting mode*. Fluorescence intensities, in this case, are expressed as the number of photons detected in a given time interval. The photocathodes of these photomultipliers may need to be cooled in order to reduce thermal emission that manifests as noise. For emission wavelengths in the red part of the spectrum, semiconductor photodiodes can be used. The multiplex photometers shown in Figure 7.8 using diode arrays or CCDs can equally be used for fluorescence studies, provided the emission intensity is high enough.

7.3.3 Fluorescence Studies using Intrinsic and Extrinsic Probes

Fluorescence spectroscopy of proteins that lack fluorescent cofactors is limited to tryptophan fluorescence. For proteins that contain fluorescent cofactors such as flavins or chlorophylls, the variety of fluorescence experiments that use these cofactors as a probe is huge. The emission of the cofactor can convey information on its molecular environment, on protonation states, or on conformational changes. Tetrapyrrole structures like chlorophylls are strongly fluorescent; it is self-evident to characterize chlorophylls and chlorophyll–protein complexes by fluorescence studies. For chlorophylls, the quantum efficiency for the isolated pigment in a solvent can be almost one, meaning that for almost each absorbed phonon there is an emitted one.

Figure 7.21 shows the absorption and emission spectra of a bacteriochlorophyll-*a* containing light-harvesting complex from the photosynthetic bacterium *Rhodobacter capsulatus*. Its function is to efficiently absorb light energy and to transfer it to the photoreactive centers. The arrangement of pigments in these complexes is optimized to funnel light energy in the form of excited states through different antenna populations towards the center of the photosynthetic apparatus to reaction centers, where excited states are trapped and used to drive charge separation. The excellent overlap of the absorption spectra of the inner parts of the pigment population with emission spectra of the population further outside forms the basis of an efficient energy transfer. The pigments involved – in this case bacteriochlorophyll *a* – exhibit singlet transitions in the visible part of the spectrum and in the near-IR: $S_0 \rightarrow S_3$ at $\approx$400 nm, $S_0 \rightarrow S_2$ at $\approx$600 nm, and $S_0 \rightarrow S_1$ at $\lambda > 750$ nm. The lowest energy transition at >750 nm varies from one antenna population to the other in order to channel and funnel energy towards the center of these structures. If this energy flow is interrupted at some point, for example, by deleting genetically a population or by extracting a pigment population, energy has to be immediately released by fluorescence emission, which makes the pigment act like a lightning conductor. This is required to avoid critical states of the chlorophyll pigments where irreversible photochemical processes could occur that would damage the pigments.

Apart from using the intrinsic fluorescence of aromatic amino acid side chains or the cofactors that are eventually present, extrinsic fluorescent probes ("fluorescence labels") can be used as so-called "reporter groups." Applications are numerous: Fluorescence probes can be used to probe the fluidity of membranes, the mobility of lipid or protein domains, or diffusional/rotational parameters. Other fluorescent probes are sensitive for membrane potential and can be used as "molecular voltmeters." Some fluorescence probes are sensitive to ion concentration or pH and can be used to follow intracellular ion transport processes.

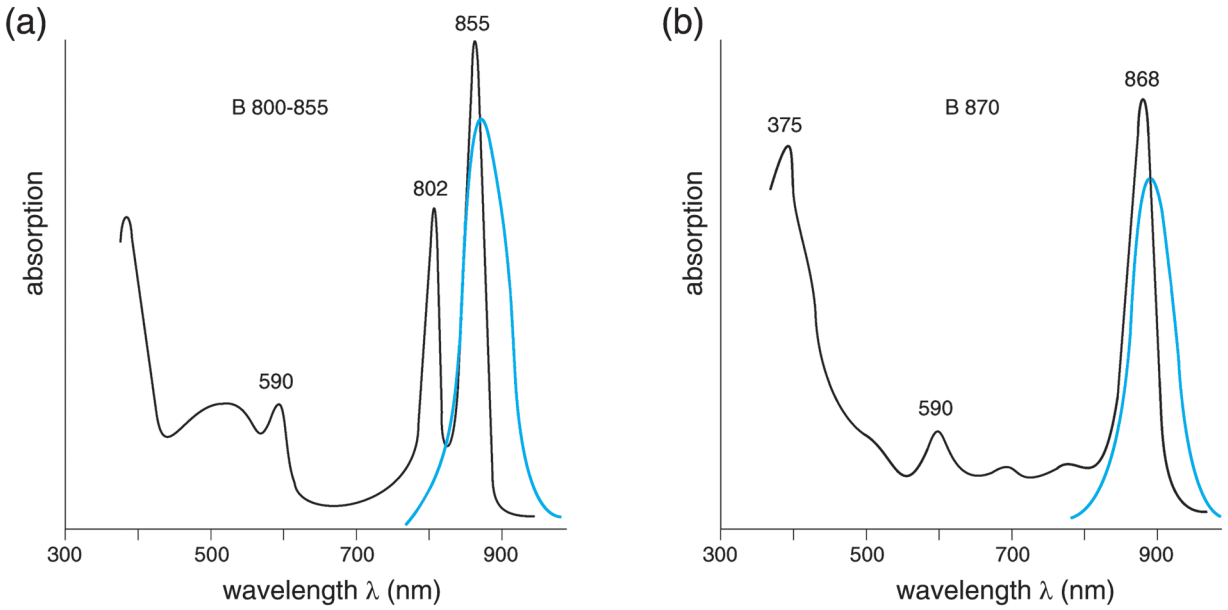


Figure 7.21 Absorption and emission spectra of light-harvesting pigments from the photosynthetic bacterium *Rhodobacter capsulatus*. (a) Absorption spectrum (black line) of the peripheral complex with two different pigment populations, one absorbing at 800 and one at 850 nm. If this complex absorbs light, rapid energy transfer occurs among the 800 and 850 nm pigments and fluorescence is only observed from the lowest energy transition of B850. The maximum of this fluorescence emission (blue line) is at approx. 865 nm. (b) Absorption spectrum (black line) of the light-harvesting complex B870 that is directly associated with the reaction center. Its absorption perfectly overlaps with the emission of the peripheral B800-85 complex.

Fluorescent probes can be bound covalently to specific amino acid side chain groups and report on the molecular environment of these amino acids. This can be made much more specific by combining site-specific mutagenesis with fluorescence labeling, eventually reducing the binding of the fluorescent probe to a single site. Other fluorescence probes bind to hydrophobic structures and can thus be integrated into the membrane lipids. They can report on the membrane potential through the position of their emission maximum and their emission intensity. The selection of fluorescent probes is huge, and many of them can be tailored to the respective application. Section 6.2 describes some of the fluorescence labels that are commonly used in bioanalysis, and that can be analyzed using routine fluorimeters.

Reagents for Introducing Fluorophores, Section 6.2

7.3.4 Green Fluorescent Protein (GFP) as a Unique Fluorescent Probe

Green fluorescent protein (GFP) and its variants have become versatile probes in molecular genetics. Some of its applications are described in Sections 8.4, 16.7.4, 34.4, and 42.2. GFP, a protein from the jellyfish *Aequorea victoria*, was discovered and characterized in the 1970s and has become one of the most important tools in molecular genetics. In this protein, composed of 238 amino acids, the chromophoric group is formed by cyclization of the side chains of three neighboring amino acids (Ser65-Tyr66-Gly67, see Figure 7.22). No additional chromophore is required.

This "auto-chromophore" exhibits two intense absorption bands with maxima at 396 and 475 nm. The shorter wavelength maximum arises from the protonated form, the longer wavlength one from the unprotonated form; the bands reflect a protonation equilibrium. The photophysics of GFP suggests emission at 459 nm upon excitation at 396 nm based on ultrafast measurements; however, emission at 508 nm – intense green light – is observed independent of the excitation. This can be explained by an ultrafast deprotonation upon absorption at 396 nm and a shift of the equilibrium to 475 nm, thus resulting in emission exclusively from the transition of the chromophore form shown in the lower left-hand side of Figure 7.22.

The unique role of GFP is based on the fact that it can be co-expressed with proteins to produce a fluorescent probe covalently bound to a protein that can be used for expression control and

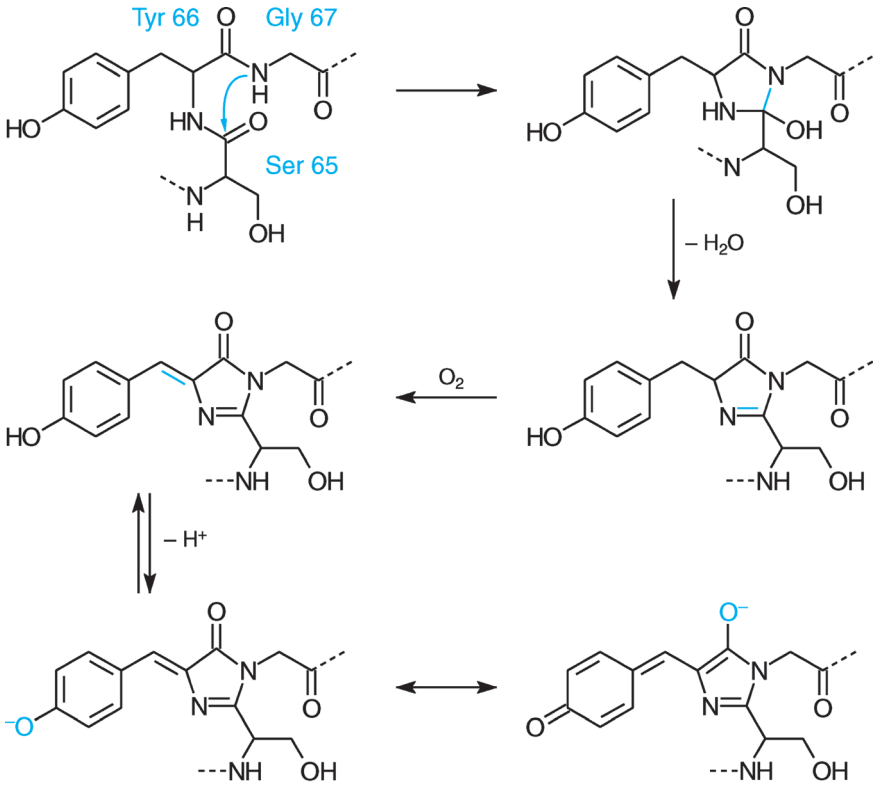


Figure 7.22 Formation of the chromophoric group of GFP from three neighboring amino acids (Ser65-Tyr66-Gly67) by cyclization and oxidation. The chromophore exists in two forms in a protonation equilibrium shown on the lower left-hand side. Taken from Steipe, B. and Skerra, A. (1997) Biospektrum, **3** (1), 28–30.

localization of the protein in cells, organelles, and tissues by fluorescence microscopy. No additional fluorescence label is required, and the tedious procedure of covalently linking a fluorescence label to an amino acid side chain is obsolete.

This relative ease of use and its unusually high fluorescence quantum yield have made GFP a "pet" of molecular biologists for the analysis of complex expression and assembly processes in biology. There are now are variants of GFP available in the form of their DNA sequence to be added to target protein sequences. These variants exhibit different absorption and emission maxima and are called, for example, BFP, CFP, or YFP for their respective emission in the blue, cyan, or yellow part of the spectrum.

Notably, however, GFP and its variants are large units that can be the size of the target protein itself or even larger, and, even if attached by a flexible linker chain of amino acids, may impact the dynamics or interactions of the target protein. This limitation will be discussed in Section 7.3.8 on the possible errors in fluorescence spectroscopy.

7.3.5 Quantum Dots as Fluorescence Labels

Quantum dots (QDs) are nanoscale semiconductor structures made, for example, from CdSe or InGaAs. Their absorption and emission properties are determined by the spatial confinement of the charge carriers in all three dimensions. Quantum mechanically they can be described as electrons that move in narrow potential wells; these electrons exhibit discrete energy levels that depend on the width of the well, that is, on the dimensions of the semiconductor structure, which typically consists of 10^4 atoms. QDs can be prepared by wet chemical techniques, by molecular beam epitaxy, etching, or by electron beam lithography. Although somewhat dependent on the type of semiconductor, a QD emitting in the blue range of the spectrum will require a size of 1–2 nm, while for emission in the red part of the spectrum the size increases to 5–6 nm. An important factor is their homogeneity in size, since this determines the width of the absorption and emission bands. In particular, QDs prepared by wet chemical methods exhibit a broad range of size and thus show heterogeneous optical properties. Quantum dots are great fluorophores that do not exhibit bleaching and that can be custom tailored to the required

wavelength, but they are heavy structures compared to protein or DNA molecules and, thus, hamper mobility.

7.3.6 Special Fluorescence Techniques: FRAP, FLIM, FCS, TIRF

Based on fluorescence measurements, numerous techniques methods have been established that use the sensitivity of fluorophores to their environment or the lifetime of fluorescence. We will discuss only some of them briefly.

Fluorescence Recovery after Photobleaching (FRAP) This technique can be used to study the properties of rotational or translational diffusion of biological molecules in membranes. In a first step, fluorescent probes are coupled to membrane lipids and the fluorescence is observed in the focus of a microscope. In a second step, a strong laser pulse is focused on the lipid membrane in order to irreversibly bleach and destroy the fluorescence dye in the focus. As a consequence of this "photobleaching," a black, non-fluorescing spot is observed in the microscope image. This spot is gradually "filled" by diffusion with unbleached, fluorescent dye molecules, leading to recovery of the original fluorescence. The time constant for fluorescence recovery is related to the lateral diffusion properties of the lipid molecules and yields information on membrane fluidity. For the measurement of rotational diffusion, selective bleaching with polarized light is used and recovery of fluorescence by rotation is measured. Both techniques rely on the fact that the lifetime of fluorescence is of the order of the rotational diffusion time, that is, around 10^{-9}–10^{-7} s.

Fluorescence Lifetime Imaging (FLIM) This technique is frequently used for imaging in cell biology. It makes use of the analysis of fluorescence lifetime, which depends on competing deactivation processes. These competing processes, for example, radiationless deactivation, singlet-triplet transfer, or photochemical reactions, depend on the local environment. Imaging is based on these environment-specific fluorescence lifetimes.

Fluorescence Correlation Spectroscopy (FCS) This method can be used to analyze dynamic processes in very small volumes and provides access to single molecule measurements (see the section below). The fluorescent molecules are analyzed in the focus of a microscope with measuring volumes significantly below 1 μm^3. Provided the fluorescent molecules in this nano-volume are sufficiently dilute, single molecules can be observed. Fluorescence photons from this volume that are counted as a function of time are subjected to fluctuations of the molecules and thus to Brownian movement. FCS is used to determine diffusion constants or reaction rates in photophysical and photochemical reactions of single molecules.

Total Internal Reflection Microscopy (TIRF) This technique uses the evanescent wave that is observed at the boundary surface upon total reflection to detect the fluorescence of molecules close to this surface, without disturbing overlap of fluorescence from the bulk volume. Cells are brought onto the surface of an internal reflection element (IRE, see also Section 7.4.3) to observe processes in the cell wall in the very short range of the evanescent electromagnetic wave. There are numerous examples of the analysis of living cells by this method.

7.3.7 Förster Resonance Energy Transfer (FRET)

The quenching of fluorescence as a consequence of collision or complex formation of fluorophores has been introduced above. A special form of quenching is the direct, resonant, and radiationless transfer of energy between two molecules that are very close to each other. The term "resonance energy transfer" was coined by the physicist Theodor Förster in 1948 and described as *Förster-resonance energy transfer* (FRET, sometimes misleadingly read as fluorescence resonance energy transfer). For FRET to take place, the distance between the donor D and the acceptor A must be extremely small, typically below 10 nm. If the distance is even smaller, direct energy transfer via electron exchange takes place because of the overlap of orbitals; this is termed Dexter energy transfer. The efficiency of FRET further depends on the overlap of the emission spectrum of the donor and the absorption spectrum of the acceptor, on the relative orientation of the transition dipole moments of D and A, and on the dielectric

FRET, Section 16.7

properties of the medium between D and A. The need for an overlap of emission and absorption spectra is an energetic requirement and it may be the reason why FRET is sometimes associated with fluorescence. However, no radiation transfer takes place with FRET, but it is a resonance effect between the transition dipoles.

The rate of FRET is proportional to $(R/R_0)^{-6}$ and thus falls off rapidly with distance; R_0 is called the "Förster radius" and exhibits values between approx. 1 and 5 nm. For $R = R_0$, half of the fluorescence is quenched by FRET. Because of the strong dependence on the distance, FRET techniques are frequently called the "molecular ruler." However, this is something of an exaggeration in view of the frequently unknown orientation of D and A (which may also change during a reaction) and the properties of the medium between D and A. Molecular distances "measured" by FRET should thus be handled with care.

Meanwhile, there is a huge selection of donor–acceptor combinations of dye molecules ("FRET-pairs") with a broad spectrum of absorption and emission wavelengths that are used for FRET measurements in bioanalysis. Many of them are small organic molecules that can be bound covalently to amino acids. Others are GFP and its variants that can be co-expressed with a protein. A third group are the lanthanides (europium, samarium, terbium, dysprosium) that can be inserted into biopolymers as fluorescent atoms. Sometimes a FRET pair consists of a fluorescent donor combined with a quenching molecule as acceptor A. This combination has the consequence that, after excitation of the donor with light, energy transfer to the nearby acceptor occurs without emission from the acceptor, unless the latter is spatially separated from the donor.

A vast number of FRET investigations have shown the broad advantage of this method. In the analysis of protein folding, the vicinity of two amino acids can be demonstrated by labeling the respective amino acid with a FRET donor and a FRET acceptor. FRET can be used to verify protein–protein interaction and to follow large protein conformational changes. A rather clever application of FRET is in DNA hybridization by labeling a nucleotide sequence with a FRET pair consisting of a fluorescent donor and a quencher as an acceptor. The free nucleotide sequence in solution, because of quenching, does not exhibit light emission. However, as soon as it is bound to a target sequence, the donor and the (quenching) acceptor become separated and fluorescence of the donor is observed. This technique is known as "blinker" or "molecular beacon."

Despite the great success of FRET techniques in bioanalysis, it should not be forgotten that the binding of FRET probes always represents an invasion into a biopolymer that can alter structure, conformation, reactivity, and dynamics. This is particularly the case for co-expressed GFP and its variants, which, because of their size, can impact proteins sterically and dynamically.

7.3.8 Frequent Mistakes in Fluorescence Spectroscopy: "The Seven Sins of Fluorescence Measurements"

A large number of fluorescence measurements describe emission studies performed to analyze the interaction between two molecules, in most cases between a protein molecule and a small molecule that is used as fluorescent reporter group. Typically for these studies, the intensity of fluorescence is monitored as a function of concentration, and the decrease of intensity is then interpreted as arising from FRET processes due to close binding. Sometimes, binding constants are derived from these measurements. Other studies use fluorescence labels to study protein dynamics or folding by FRET techniques. However, before such fluorescence data can be interpreted seriously, one should make sure that all the possible pitfalls of fluorescence measurements can safely be excluded.

- **Sin 1: Incomplete or wrong labeling with fluorophores**

  With the exception of intrinsic protein fluorescence from tryptophan, fluorescent labels have to be added to the biological polymer, either by covalent binding to an amino acid, by hydrophobic/hydrophilic interaction, or by intercalation (e.g., into DNA strands). The labeling efficiency greatly varies with the procedure, and it should be guaranteed that incomplete labeling or false labeling (at the wrong site) can be excluded. Common labeling at lysine residues is essentially not site-specific, in contrast to cysteine labeling. The aim should be to knock-out mutants that carry only one such residue.

- **Sin 2: FRET overdone:**
  Förster resonance energy transfer (FRET) is a great technique as long as the rules are followed. It has been greatly promoted as the "molecular ruler", and its application does not require expensive instrumentation. However, FRET efficiency strongly depends on several parameters that have been discussed above and that are difficult to control. If all the variables (relative orientation of donor and acceptor, dielectric properties of the medium between them, electronic levels of donor and acceptor) are well controlled, FRET is an elegant tool by which to measure distances between two sites in a biopolymer. However, many scientists show blind trust in their experiment and report distances to fractions of an ångstrøm, although the orientation of the donor–acceptor couple and the medium between them is unknown and may vary during the experiment, in particular if FRET is used to study distance changes in the course of a conformational change or protein folding. The fact that donor–acceptor FRET couples are offered for a certain distance range is somehow seductive, but scientists should be aware that FRET is still more of a rather coarse estimation of a distance than a precision tool that is better than crystallographic data. Usually, the orientation and mobility of the labels is a big "unknown" and cannot be assumed to be fixed or perfectly random.
- **Sin 3: GFP overdone:**
  GFP is an excellent fluorescence label because of its easy attachment to a protein. GFP and its variants are unsurpassed when it comes to the task of demonstrating protein expression, localization, or translocation in cell biology. There are, though, many studies on proteins and enzymes that use GFP and its analogs for folding, protein dynamics, docking, and other functional analysis. However, the protein studied is sometimes comparable in molecular mass to GFP (27 kDa) and its attachment to a protein may impair the dynamics severely. In some studies where folding was analyzed with GFP labels the molecular weight of the two GFP-type labels used for FRET by far exceeded the molecular weight of the protein under investigation. This is like an attempt to study the mobility of a dog by adding heavy weights at the tail and at the head.
- **Sin 4: Quantum dots overdone:**
  Quantum dots (QDs) are quite popular in fluorescence microscopy because they are not prone to bleaching in the way that many organic fluorescence dyes are and because they can be prepared for almost all wavelengths. However, the user of QDs must be aware that they are nanoscale objects that seriously hamper dynamic processes of proteins and enzymes. Similar to GFP, QDs are great tools to label structures in cell biology, but their use for the study of molecular and dynamic properties should be performed much more carefully.
- **Sin 5: Shading and Inner Filter effect I:**
  In many cases where fluorescence studies are used for the analysis of binding or docking of a smaller molecule X to a protein Y, the absorbance of the substance X either for the excitation light or for the emitted fluorescence light is neglected. As a trivial effect, addition of substance X to protein Y then leads to additional absorbance of the exciting light, or to absorbance of emitted fluorescence. These effects are called shading and inner filtering; in both cases the fluorescence intensity at the detector is decreased. Dilution of the substances X and Y can help; if not, the decrease of fluorescence intensity can be misinterpreted as quenching induced by binding.
- **Sin 6: Shading and Inner filter effect II:**
  In many published fluorescence studies on ligand binding to a protein the protein absorbance at 280 nm, resulting from aromatic amino acid side chains, shows total values of >1, sometimes even 2–3. If tryptophan emission from these intrinsic fluorophores is used for analytical purposes under these conditions the authors should be aware that the emission intensity measured is a delicate balance between the UV light at approx. 280 nm that reaches the aromatic amino acids and the emitted light at 300–350 nm that reaches the detector.
- **Sin 7: Fluorescence of complexes:**
  The binding of a small molecule as a ligand to an enzyme or a protein can lead to reduction of the fluorescence intensity by quenching, assuming that the protein–ligand complex is non-fluorescent. The term "apparent quenching" would be more appropriate here, because the complex formed may exhibit an altered fluorescence. If the fluorescence intensity is used to determine a binding constant, the (eventually altered) complex fluorescence can be misinterpreted and may result in a wrong binding constant.

In summary, many fluorescence studies suffer from one of these sins and need critical assessment. The relative ease of fluorescence measurements is tempting; however, inappropriate methodologies should be avoided.

7.4 Infrared Spectroscopy

7.4.1 Basic Principles of IR Spectroscopy

The infrared spectral region is adjacent to the visible part of the spectrum and is known as the *near-infrared* (NIR, ≈ 750–3000 nm), *mid-infrared* (MIR, ≈3–30 µm), and *far-infrared* (FIR, 30–1000 µm) (see also Table 7.1). The terahertz (THz) region, which has also been explored recently for bioanalytical and biomedical spectroscopy, includes the FIR region but also extends to the microwave range; one of its applications is the "body scanner" for security control on airports. In IR and Raman spectroscopy it is common to use wavenumbers (the number of waves per cm) instead of the wavelength λ.

The dominant transitions in the mid-infrared are among vibrational levels, as indicated in the term scheme in Figure 7.6. Rotational transitions are found in the far-infrared. The near-infrared part of the spectrum shows some overtones of vibrational MIR transitions as well as very low-energy electronic transitions.

To understand these phenomena, we return to the model of the diatomic molecule of two masses elastically connected by a spring, which we have already used in Section 7.1. The atomic masses m_1 and m_2 can move relative to each other by extending and compressing the spring from its equilibrium distance r_0 (Figure 7.23). In classical mechanics, this system has the potential energy $E = ½k(r - r_0)^2$, where $(r - r_0)$ is the deviation from the equilibrium distance and k is a constant describing the stiffness of the spring.

To describe the movement of two masses relative to their center of gravity (this needs more calculation effort because of the higher number of parameters) the "reduced mass" μ is introduced:

$$\mu = \frac{m_1 m_2}{m_1 + m_2} \tag{7.11}$$

The movement of this reduced mass against a fixed position (Figure 7.23) can be described by a potential that exhibits a parabola form (a "harmonic potential") with its minimum at r_0 (Figure 7.23b). This potential yields the vibration frequency υ:

$$\nu = \frac{1}{2\pi} \cdot \sqrt{\frac{k}{\mu}} \tag{7.12}$$

This model is called the harmonic oscillator. While the classical description of the harmonic oscillator can assume any energy level, its quantum mechanical treatment requires discrete

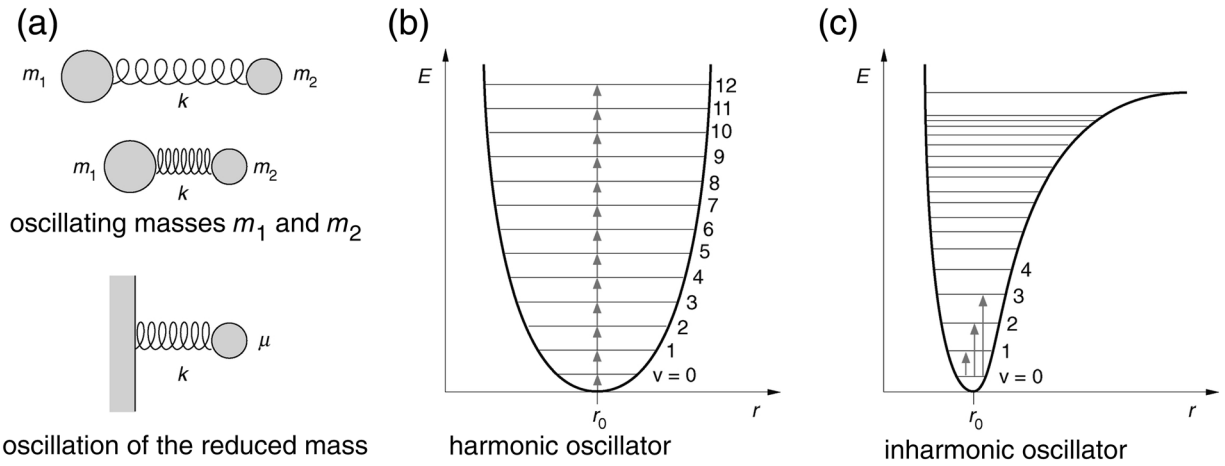


Figure 7.23 Vibration properties of a diatomic molecule (a) and their description with harmonic (b) and inharmonic (c) oscillators.

energy levels at equal distance characterized by vibration quantum numbers $v = 0$, 1, 2, 3, and so on:

$$E(v) = \left(v + \tfrac{1}{2}\right) h\nu_{vib} \tag{7.13}$$

where h is the Planck constant (quantum of action). For $v = 0$, the quantum mechanical harmonic oscillator yields the zero-point energy $E = \frac{1}{2}h\upsilon_{vib}$ as a consequence of the Heisenberg uncertainty relation. Furthermore, the selection rules allow only transitions to the neighboring level, that is, $\Delta v = \pm 1$.

7.4.2 Molecular Vibrations

For real molecules, the harmonic oscillator model cannot be applied because the electrostatic repulsion of the nuclei leads to a strong increase of the potential energy upon reduction of the distance, much steeper than given by a harmonic potential. Furthermore, a large increase of the distance of the nuclei leads to a dissociation of the molecule that is not reflected by the harmonic potential. This leads to the inharmonic oscillator model where the potential energy rises steeply if the nuclei approach, and where an increase in the distance leads to dissociation (Figure 7.23c). The energy levels are no longer equidistant and selection rules allow transitions to higher vibration levels ($\Delta v = \pm 1, \pm 2, \pm 3, \ldots$). These transitions can lead to overtones and combinations of vibrations and explain the absorption properties of molecular vibrations that can range into the near-infrared and even to the visible part of the spectrum, even though the theoretically highest possible fundamental vibration, that of the hydrogen molecule H–H (although not infrared active), is in the mid-IR. We are all familiar with higher overtones of the water H–O–H vibration that absorbs in the NIR around 1.6 µm and at 900 nm and that are responsible for the absorption of red and far-red light by thick layers of water (e.g., in a swimming pool).

Molecules can also perform rotations in addition to vibrations. In the model of the diatomic molecule these rotations occur, for example, around an axis perpendicular to the line connecting them. The rotational frequencies are much lower than the vibrational frequencies; consequently, these lower-energy transitions lead to absorbance in the far-Infrared. Rotations couple with vibrations and lead to rotational substructures of vibrational spectra, a phenomenon that is observed in the IR spectroscopy of gases. In biopolymers, this cannot be detected because of heterogeneous broadening and thus wider bands.

If we want to use these basic principles to describe polyatomic molecules, we first have to define the possible movements of the molecules and individual atoms. To define the position in space, the three coordinates x, y, and z are required, which adds to $3N$ coordinates for N atoms, so-called "degrees of freedom". We can reduce these $3N$ degrees of freedom by three degrees of freedom for the translational movement of the entire molecule and three for the rotation (two for the rotation of a linear molecule); we thus keep $3N$–6 degrees of freedom for a nonlinear molecule and $3N$–5 for a linear molecule. These are the normal vibrations or normal modes of the molecule. The normal modes of CO_2 and H_2O molecules are shown in Figure 7.24 and their designations are also given in Table 7.3.

The existence of a normal mode is only one of the prerequisites for a molecule to absorb energy from the electromagnetic wave with a suitable frequency and to undergo a transition to a higher vibrational level. In addition, the dipole moment of the molecule must also change with the normal mode. If this is the case, the oscillating electric field vector can interact with the molecular dipole if the frequencies match. The dipole strength governs the probability of the absorption process and, thus, the intensity of the absorption band.

Absorption arising from vibrational transitions is generally weaker than absorption due to electronic transitions. "Highly allowed" (i.e., strong) electronic transitions can reach molar absorption coefficients higher than 10^5 (Table 7.2). Absorption coefficients for vibrational transitions are in the range below 10^3. As an example of a rather strong IR absorber, we consider the IR absorption of the carboxyl R-COOH group from aspartic or glutamic acid side chain groups. The C═O bond of the protonated form absorbs in the range 1700–1750 cm^{-1} depending on H bonding and polarity, and exhibits molar absorption coefficients between 100 and 300 l mol^{-1} cm^{-1}. Another example is the C═C of the aromatic ring of the tyrosine side chain located around 1500 cm^{-1} with similar absorption strength.

Table 7.3 Normal modes of CO_2 and H_2O.

	Vibration mode	Abbreviation	IR activity	Wavenumber (cm^{-1})
CO_2	Symmetric stretching mode	υ_s	No	
	Antisymmetric stretching mode	υ_{as}	Yes	≈2200
	Bending mode (in plane/out of plane)	δ	Yes	
H_2O	Symmetric stretching mode	υ_s	Yes	≈3350
	Antisymmetric stretching mode	υ_{as}	Yes	≈3300
	Symmetric bending mode	δ_s	Yes	≈1660

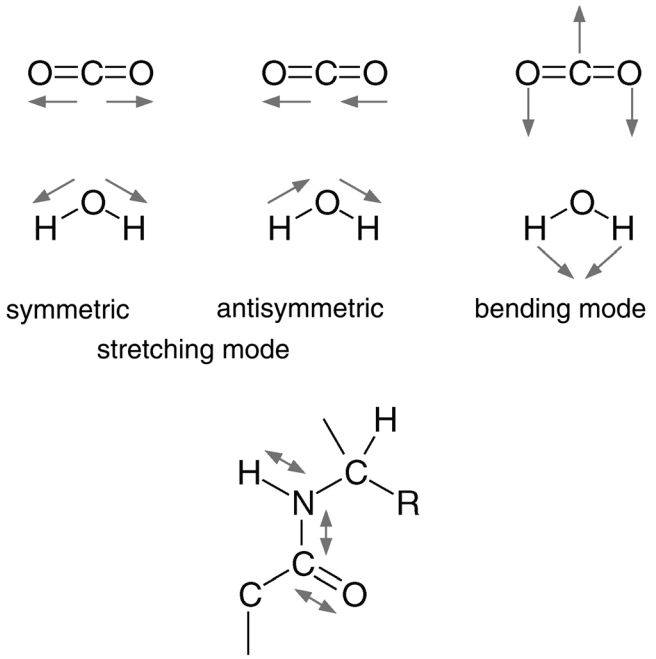


Figure 7.24 Normal modes of the CO_2 and the H_2O molecule. The arrows indicate the alteration of bond lengths and bond angles. For the amide bond, only the most relevant modes are labeled.

The O—H bending mode of water has a rather small absorbance coefficient of $<20\,l\,mol^{-1}\,cm^{-1}$, but the fact that the water concentration (in water) is approximately $55\,mol\,l^{-1}$ leads to a strong background of water in aqueous samples. Water absorbance in the range of the O—H bending mode around $1660\,cm^{-1}$ is about 1 for an ultrathin layer 10 µm thick. As we shall see below, recording IR spectra of aqueous samples thus requires ultrathin layer cuvettes.

7.4.3 Technical aspects of Infrared Spectroscopy

To measure the infrared spectra of biopolymers we can use double beam spectrophotometers as already described in Section 7.1.4 for the visible part of the spectrum or the UV. The IR spectral region that contains bands of diagnostic value ranges from approx. 5000 to $500\,cm^{-1}$, corresponding to wavelengths between 2 and 20 µm. The selection of appropriate optical components, light sources, and detectors, however, deserves special attention. Lenses for an IR setup can only be made from materials that are transparent in that range (or in part of it), for example, from NaCl, KBr, CaF_2, ZnS, ZnSe, Ge, Si, or diamond. Some of these materials are hygroscopic, others exhibit a high refractive index and thus high reflectivity, and materials such as sapphire or diamond are quite costly. To avoid this, IR setups use almost exclusively spherical, elliptical, or parabolic mirrors for the measuring beam optics. Similar considerations are required for the materials for IR cuvettes. For aqueous samples, because of water background absorption, the path length should be 20 µm or lower (about the thickness of a human

(a)

material	permeability	characteristics
CaF_2	190 nm – 10 µm	universally applicable for UV/VIS/IR, very hard
BaF_2	200 nm – 12 µm	universally applicable for UV/VIS/IR, brittle
ZnSe	500 nm – 20 µm	highly-refractive, suitable for ATR crystals
ZnS	450 nm – 15 µm	highly-refractive, suitable for ATR crystals
Ge	1,2 µm – 25 µm	highly-refractive, needs an anti-reflective layer
Si	1,2 µm – 20 µm	highly-refractive, needs an anti-reflective layer

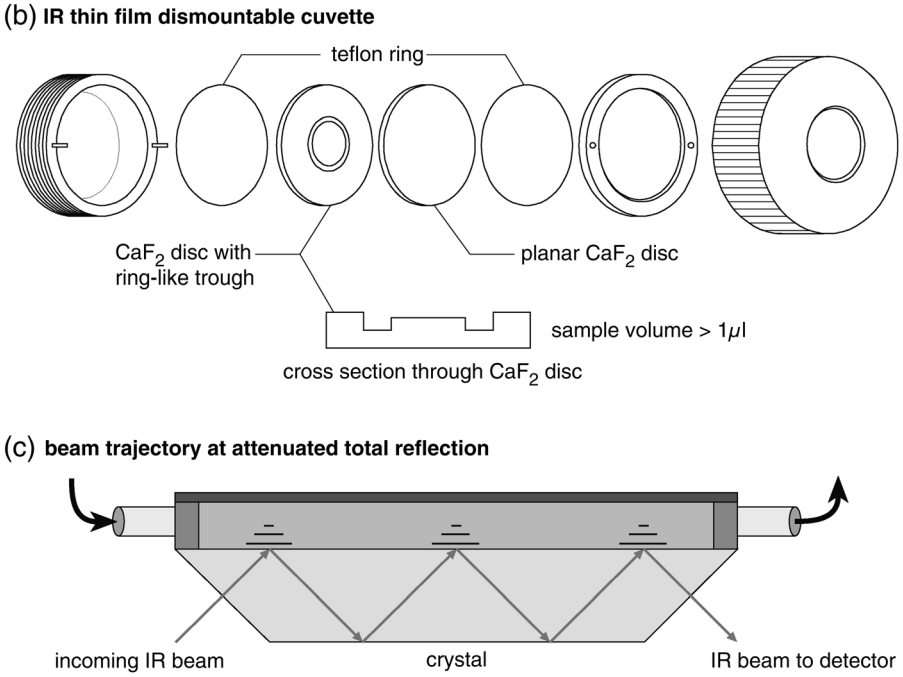


Figure 7.25 (a) Typical IR window materials and their transmission. Only water insoluble and broadly transparent materials have been listed. (b) Exploded view of an IR micro-cuvette that can be disassembled. The window shown in cross section has a trough in the center that defines the path length. It is surrounded by a second ring-like trough that takes excess sample material. This window is covered with a second, flat window to form a cuvette with highly reproducible path lengths from <5 µm to >50 µm. (c) Schematic representation of an ATR experiment. The IR beam enters a highly refractive material (e.g., Ge, ZnSe, ZnS, see part (a)) and propagates in this light guide. At the interface where total reflection occurs, an evanescent wave emerges and measures the properties of the molecules dissolved or adsorbed. The penetration depth depends on the refractive index, the entry angle, and the wavelength, and is of the order of half the wavelength. The ATR cell shown here is designed as a flow cell that can be coupled with, for example, chromatographic methods.

hair), which is much shorter than for UV or VIS spectroscopy. A consequence of these path lengths is the need for highly concentrated samples because of the low absorbance coefficients discussed above.

Among the possible window materials, CaF_2 is a favorite because it is water-insoluble and exhibits full transmission from the UV (190 nm) to the mid-IR (≈10 µm). Glass becomes nontransparent already in the near-IR, and quartz can only be used to about 5 µm (2000 cm^{-1}). Many other possible window materials can be discarded because of their water solubility (NaCl, KBr), high reflectivity (Ge, Si, ZnSe, ZnS), because they are highly toxic (thallium compounds), or have a high price (sapphire, diamond), although some of these materials may be useful for special applications. In the case of high reflectivity, surface coating can improve the usability.

A cuvette system that has proven to be suitable for quantitative IR spectroscopy of biopolymer samples in water is shown in Figure 7.25. It allows reproducible path lengths from <5 µm to >50 µm and can be disassembled for cleaning and filling. Using CaF_2 as window material, the spectral range is up to 10 µm, which is sufficient for most applications. The cell is filled by pipetting some µl of the sample onto a trough-shaped window that is then covered with a flat window. These cells have almost no dead volume and require only 1–5 µl sample volume, depending on the path length.

An alternative to transmission measurements is measurement by attenuated total reflection (ATR). For ATR, an infrared transparent light guide made from highly refractive index material is used ("internal reflection element", IRE). It is shaped to cause an IR beam that enters from the side to be totally reflected at the IRE surface. The number of total reflections is determined by the length of the element, the refractive index, and the angle that is used to couple the IR beam into the element; 1–10 reflections are common.

Internal total reflection causes a standing electromagnetic wave formed by interference of the incoming and the reflected beam. As a consequence of the standing wave inside the IRE, an evanescent wave forms outside the IRE for reasons of continuity of the wave. This evanescent wave penetrates the material in contact with the IRE and contains information on the IR absorption of the sample. The depth of penetration depends on the wavelength, the entrance angle, and the refractive index, and is typically 0.5–1 µm, a fraction of the wavelength per total reflection. The advantage of this technique is that the effective path length is defined by the total reflection process itself, independent of the thickness of the layer above (it could even be the Pacific Ocean). This technique was originally developed for the analysis of paints and pastes that could be smeared onto the surface; in bioanalytical IR spectroscopy it has become very popular because of the simple sample access with this optical interface. Practical applications include the analysis of body fluids, tissues, or fermentation products (e.g., for quality analysis of beer or dairy products).

Apart from window materials, light sources have to be adapted for IR spectroscopy. In spectrometers, typically ceramic materials are used that are electrically heated to 700–1000 °C. They are called globars (from *glo*wing *bar*) and emit, to a good approximation, blackbody radiation with a maximum in the NIR and a small fraction of visible light, which gives a convenient way to adjust the optics. Care should be taken that this MIR and VIS radiation is blocked before reaching the sample in order to avoid heating, for example, by using a germanium window that cuts all wavelengths <1600 nm.

Special applications use tunable IR lasers, such as the CO and the CO_2 laser, which are powerful but with only small tuning ranges. Semiconductor lasers are available from europium doped PbS or PbSe semiconductors; however, they work only at cryogenic temperatures and are thus difficult to use. A new class of intersubband lasers called quantum cascade lasers (QCLs) is very promising for IR spectroscopy. Charge recombination in QCL occurs in a cascade of insulator/semiconductor layers with high efficiency for IR photon emission. These lasers are available as DFB (distributed feedback) lasers emitting with high power at fixed IR wavelengths at almost any part of the spectrum. Meanwhile, these quantum cascade lasers are available with external resonators (external cavity QCL, EC-QCL) and can be tuned by over up to 400 cm^{-1}. These easy-to-use and intense IR sources will most probably revolutionize bioanalysis, in scientific as well as routine sensor applications.

Detectors for IR can be thermal detectors such as thermocouples or pyroelectric detectors; both convert the heat from absorption of IR radiation into an electric signal. While they are used for broad-band detection in routine spectrometers, research applications in IR spectroscopy use quantum detectors instead. They use the internal photoelectric effect in semiconductors such as indium antimonide (InSb) or mercury cadmium telluride (HgCdTe) and yield either a change of photoconductivity or a photocurrent proportional to the intensity. The sensitivity of these detectors is strongly dependent on the wavelength and the temperature. Cooling (either thermoelectrically or with liquid nitrogen) is thus mandatory to obtain low noise, high sensitivity, and high time resolution.

We have discussed above the disadvantage of dispersive techniques where all spectral elements are successively scanned with a monochromator. This is even more relevant for IR spectroscopy, because the photon density emitted from thermal sources is low, and the photon energy is low, slightly above thermal energy at room temperature (for IR laser sources, the photon flux is much higher). The spectral power of a typical thermal IR source in the mid-IR at a spectral resolution of 2–4 cm^{-1} is at most a few microwatt.

The multiplex methods that we have discussed for the UV/VIS range have not entered routine applications. One-dimensional or two-dimensional infrared detector arrays are still quite expensive and – at least that for high sensitivity and resolution – have been limited to military applications for thermal detection.

Instead of using multi-detector arrays, another multiplex method has been established for routine IR spectroscopy that allows simultaneous recording of many wavelengths by the use of an interferometric technique. This multiplex method is called Fourier transform infrared (FT-IR) spectroscopy. Its principle is shown in Figure 7.26.

Fourier-Transformation, Section 18.1

Instead of directly recording a spectrum, an interferogram $I(x)$ is measured by recording the intensity I at the detector as a function of the position x of a movable mirror in a Michelson interferometer. This interferogram is the Fourier transform of the spectrum $I(\upsilon)$ that is obtained by inverse Fourier transformation. Because this transformation $I(x) \rightarrow I(\upsilon)$ requires less than a

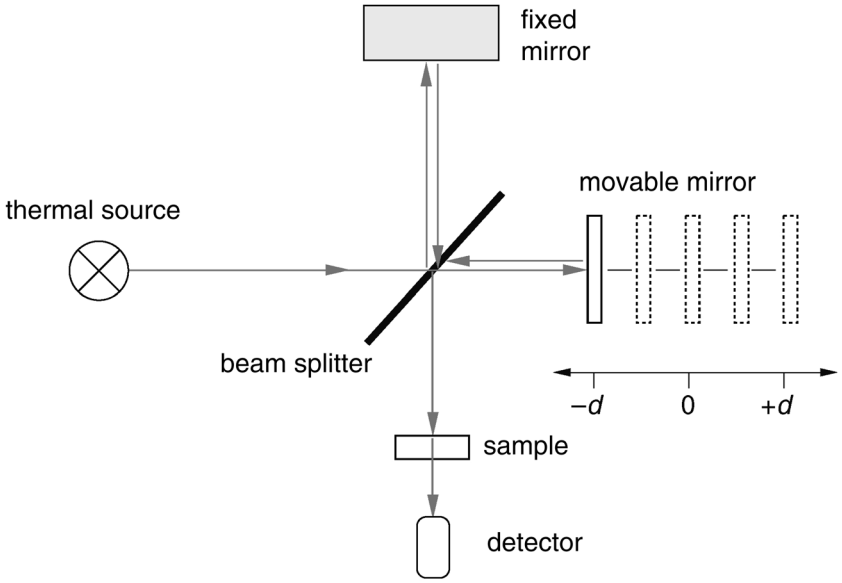


Figure 7.26 Basic principle of a FT-IR spectrophotometer. Polychromatic IR radiation from a thermal source enters a Michelson-type interferometer. There it is separated equally into a transmitted and a reflected beam by a beam splitter. The part reflected by the fixed mirror is recombined with that reflected from the movable mirror. Depending on the difference in path length, constructive and destructive interference occurs. As the movable mirror travels from $-d$ to $+d$, an interference pattern is generated depending on the wavelength and phase difference. This recombined, interference modulated beam passes through the sample and reaches the detector. The intensity recorded as a function of the position x from $-d$ to $+d$ is the interferogram $I(x)$.

second with modern processors, the spectra are instantaneously obtained. It was this innovation in the 1970s that made possible many of the present routine bioanalytical applications of infrared spectroscopy.

FT-IR spectroscopy as a multiplex technique facilitates the recording of infrared spectra of biopolymers at a time resolution from milliseconds to nanoseconds using specialized techniques – a key to functional investigations on proteins. Even for simple routine FT-IR instruments, the movement of the mirror takes less than a second. It takes no further effort or equipment to record spectra in these time frames. In cases where time resolution is not required, interferograms can be recorded and averaged before calculation of the spectrum by Fourier transform. This results in excellent signal-to-noise ratio for FT-IR spectra of proteins, at a noise level much smaller than the absorbance contributions from individual bonds in a huge biopolymer. We will see below that, upon using difference techniques and perturbation techniques, these contributions can be isolated and interpreted in terms of molecular details of reaction mechanisms.

Special "rapid scan" forms of interferometers have been built that are optimized for rapid mirror movement and data acquisition. The measuring time for an interferogram can be reduced to some milliseconds by recording one for the forward and one for the backward movement of the mirror – in principle they should be identical. For even faster time resolution to nanoseconds, the step scan technique is used. However, this requires the repetitive excitation of the sample (between 10^4 and 10^5 times) in identical form and has been successful for only a few photobiological systems.

7.4.4 Infrared Spectra of Proteins

The concept of normal modes of a molecule discussed above may work for small molecules but it is difficult to use even for medium sized molecules and for proteins. Even a relatively small protein with a molecular mass of 12 kDa (e.g., a soluble *c*-type cytochrome) consists of approx. 100 amino acids and already shows hundreds of normal modes. For these larger molecules, it is more convenient to use the concept of group vibrations that is used in chemistry for the identification of compounds by infrared spectroscopy. This concept is based on the following assumptions: The macromolecule is formally dissected into individual bonds and groups that, to a first approximation, each vibrate independently. Each group or bond has its individual vibration pattern that depends on force constants and atomic masses. These vibrations can be found as absorption bands in the infrared spectrum in a range specific for this group. Within that range, the detailed absorbance then depends on the ligand, that is, on the binding to the rest of the molecule. For the entire macromolecule, the spectrum is the sum of the absorptions of all groups.

Table 7.4 Vibrational modes of the peptide bond.

Symmetry	Name	absorption maximum at: (cm^{-1})	Composition[a)]
In the peptide plane	Amide A	≈3300	NH_s (100%)
	Amide B	≈3100	NH_s (100%)
	Amide I	≈1650	CO_s (≅ 80%), CN_s, CN_d
	Amide II	≈1550	NH_{ip} (≅ 60%), $C{=}N_s$ ≅ 40%
	Amide III	≈1300	CN_s (40%), NH_{ip} (30%), CC_s (30%)
Out of the peptide plane	Amide IV	≈725	NH_{op}, CN_t
	Amide V	≈625	CO_d (40%), CO_s (30%), CNC_d
	Amide VI	≈600	CO_{op}, CN_t
	Amide VII	≈200	NH_{op}, CN_t, CO_{op}

a) s: stretching mode, d: deformation mode, t: torsion mode, ip: in-plane bending, op: out-of-plane bending. The numbers give the approximate potential energy that a specific bond contributes.

The concept of group vibrations works remarkably well for large molecules such as proteins, where the following vibrational modes are expected:

- vibrational modes from the polypeptide backbone;
- vibrations from amino acid side chains;
- vibrations from cofactors eventually present;
- vibrations from detergent, lipids, water, buffers (depending on the protein).

The polypeptide backbone with the repeating peptide bond contributes most to the infrared absorption of a protein and dominates the absorption spectrum. The normal modes of the peptide bond are summarized in Table 7.4.

Of these vibrational modes from the peptide bond, the amide-I mode that mostly originates from the peptide C═O bond (Table 7.4) can be used for the analysis of the protein secondary structure. Its frequency and intensity are sensitive for the hydrogen bonding between the ─C═O and the ─N─H group, ─C═O · · · H─N─. A strong H bond weakens the double bond character and lowers the C═O frequency. The weaker the hydrogen bond, the stronger the double bond character and the higher the C═O frequency. The hydrogen bonds formed within the secondary structure lead to an amide-I frequency characteristic for this structure. This allows a relatively simple quantification of the secondary structure elements in a polypeptide and implies the following steps:

1. recording of a high quality IR spectrum of the protein in a suitable buffer;
2. subtraction of the spectral contributions from buffer, water, and so on;
3. application of procedures to improve resolution;
4. application of procedures for band deconvolution and band synthesis;
5. evaluation of the amide I band components on the basis of standard data sets from known proteins.

These procedures have become well established and yield information on the secondary structure of proteins with relatively little effort. They can be applied to soluble and to membrane-bound proteins and provide information on proteins that have not been crystallized previously. Moreover, this information is obtained from proteins in their native aqueous environment while protein crystallography possibly monitors artificial states. Nevertheless, amide-I IR spectra should be interpreted with care. The mathematical deconvolution step of the amide-I envelope in its components is quite reliable and can be well controlled. However, the assignment of these contributions to secondary structure elements is rather more critical. One reason for this is that these secondary structure elements differ slightly in their absorption depending on the length. A major problem in the interpretation is to define the borders of secondary structure elements: Should a specific amino acid be counted as at the end of the helix or at the start of the adjacent loop?

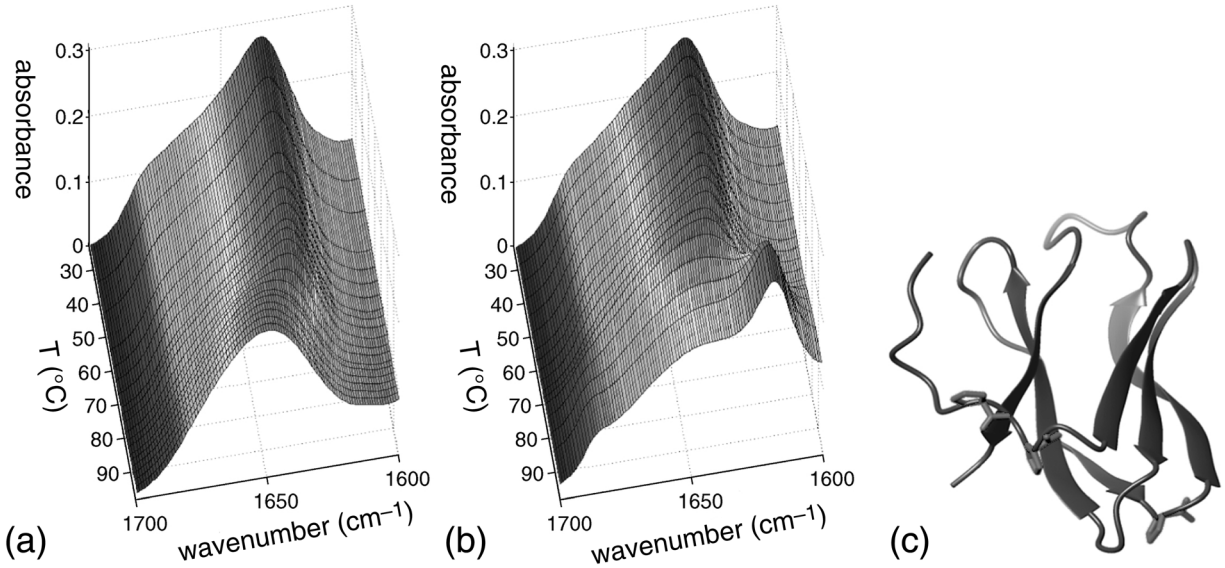


Figure 7.27 Infrared spectra of the protein tendamistat in its native form (a) and with three point mutations where three prolines were replaced by alanines (b). The temperature series was recorded after rapid heating. The spectral region displays the amide-I range with specific absorption for secondary structures. (c) Ribbon model of tendamistat with the proline residues marked.

As well as the analysis of the fractions of secondary structure elements (% α helix, % β sheet, % random coil), this method can be successfully applied to monitor the changes of these elements in the course of folding and unfolding processes as well as for the investigation of protein stability and protein denaturing. Figure 7.27 shows the IR spectra of the small soluble protein tendamistat, a known inhibitor of α amylase, upon thermally induced unfolding. The secondary structure of the protein is shown in the ribbon model (Figure 7.27c).

The native protein at room temperature (Figure 7.27a) shows the typical IR spectral pattern of a β sheet protein. Heating of the protein results in a loss of this pattern and the development of a single band characteristic for an unordered structure at temperatures above 90 °C. The "melting" β sheet structure can be directly monitored. This process is reversible: upon cooling to room temperature the pattern typical for a β sheet protein is formed again. As the spectra indicate, the folding of this protein can be described with a simple two-state model (folded/unfolded). The same protein with three point mutations (proline → alanine) at room temperature does not show any difference in structure as compared to the wild type, as demonstrated by the IR spectra. However, it aggregates irreversibly upon heating and exhibits a clear signature for large protein aggregates that is known from the spectroscopy of amyloid fibrils and plaques.

In recent decades, IR methods have been developed for the analysis of proteins that can be used to investigate the molecular details of reactions. All of them are based on difference techniques and perturbation techniques where a single sample is influenced by an external perturbation (see also Section 7.1.5). This perturbation should start the desired reaction as specifically and quantitatively as possible, and also as rapidly as possible. The initial and the final state of a reaction as well as metastable intermediates can be monitored by time-resolved IR spectroscopy and lead to a "molecular video" of the reaction in an enzyme, which can help to unravel its molecular mechanism.

All these reaction-induced difference techniques are based on the concept that difference spectra only reflect the part of the protein, the "hot spot," that undergoes changes during the reaction, and that the absorption of the "silent" parts of the protein is compensated upon difference formation. A direct calculation of difference spectra between two samples in different forms usually fails, because even small differences in the concentration or the path length would yield difference spectra with much larger amplitudes than caused by the reaction. Difference spectra generated by reaction-modulated IR difference techniques, however, allow the correlation of the disappearance of bands and the appearance of new bands or the shift of bands with the perturbation; this leads to extremely high sensitivity and specificity. Based on the sensitivity of FT-IR spectroscopy or laser IR spectroscopy, absorption changes

of 10^{-3}–10^{-5} of the total absorbance can be reliably measured and attributed to the triggering event. For a protein of a molecular mass around 100 kDa, this corresponds to the contribution of individual bonds to the total spectrum and thus guarantees individual bond sensitivity.

If the entire diagnostically relevant spectral range from >2000 to <500 cm^{-1} is utilized, a molecular picture of bond length and bond angle alterations, of protonation changes, and of conformational alterations can be collected. This data form the basis for the description of a scenario of the reaction on the level of bonds and atoms.

Originally, these IR perturbation techniques had been developed for chromoproteins where the photoreaction can be easily and specifically induced by illumination, either with a short flash or by constant illumination. This led the way to characterization of the molecular reaction steps of the visual pigment rhodopsin, the light-driven proton pump bacteriorhodopsin, and the primary reactions in photosynthesis (see also Section 7.2). Starting from these light-triggered reactions, different perturbation techniques are now available for IR difference spectroscopy that can be applied for many different proteins and enzymes:

- light-induced difference spectroscopy with laser flashes or constant illumination;
- redox-induced IR difference spectroscopy using electron transfer at transparent or reflecting electrodes in a spectroelectrochemical cell;
- photochemically-triggered difference spectroscopy with light-triggered substrates (see Section 7.1.5, Figure 7.9);
- thermally-induced difference spectroscopy with electric heating or heating with a laser flash;
- rapid diffusive mixing with thin-layer cells;
- perfusion methods with ATR flow cells.

The methodical developments of the past 30 years have made infrared spectroscopy an important and efficient technique for the characterization of biopolymers under native conditions. The reservations against IR spectroscopy that come from the early applications in chemistry, such as high (mg) sample quantities or the limitation by water, are no longer limiting obstacles. IR spectroscopy makes use of a broader spectral range than visible light spectroscopy, and can probe many more diagnostically relevant bands as compared to electronic transitions. It allows a label-free analysis of proteins. All this has increasingly led to the development of methods that yield information on structure, function, and dynamics of biopolymers. For all these new methods, which have received the well-justified name "window into the protein," we refer the reader to the Further Reading.

7.5 Raman Spectroscopy

7.5.1 Basic Principles of Raman Spectroscopy

Raman spectroscopy is related to infrared spectroscopy. It relies on a scattering effect first described by Raman and Krishnan in 1928. They observed scattered light with frequencies shifted with respect to the incoming light. This phenomenon was named the Raman Effect after one of the discoverers, and the spectroscopic method based on it is called Raman spectroscopy. Soon after its discovery this new method entered chemical analysis, but it required the development of intense laser sources before it became relevant in bioanalysis in the 1970s.

Raman spectroscopy has a respected standing for the analysis of pigmented proteins, for example, for chlorophyll-protein complexes or heme proteins, and will be discussed here immediately after IR spectroscopy because the resulting information of the vibrational spectra of biomolecules is comparable. The theory and experimental principles, however, differ completely.

To understand the Raman Effect, we consider the processes that govern the interaction of photons of an incoming light beam with the molecules of a sample. Besides light absorption, where photons of suitable energy can induce a transition to a higher electronic level, elastic and inelastic scattering can occur where the photon is not absorbed, but where its direction of movement can be altered. This is handled in the term scheme by introducing virtual levels. Elastic scattering was discussed above in connection with absorbance measurements on real sample. Raman scattering can be described as inelastic scattering of a photon at a molecule.

Upon scattering, the molecule undergoes a transition to this higher virtual level and the photon loses part of its energy.

In a classical description, the incoming light wave induces an oscillating dipole moment in the molecule that overlaps on an eventually present permanent dipole moment. Because each oscillating dipole moment generates a new electromagnetic wave, each molecule contributes its part to elastic scattering (Rayleigh scattering) and to inelastic scattering (Raman scattering).

The oscillating dipole moment $\mu(t)$ can be described as:

$$\mu(t) = \alpha(\nu)E_0 \cos 2\pi\nu t \tag{7.14}$$

where α is the polarizability tensor and E_0 is the electric field strength. In this model, α is not a constant because of the vibrations of the molecule, and thus needs to be described with its time-dependence:

$$\alpha(\nu) = \alpha_0(\nu) + \alpha'(\nu)\cos 2\pi\nu' t \tag{7.15}$$

In this equation, α_0 represents the equilibrium polarization, α' the change of the polarizability upon movement of the nuclei (ν'), and ν' is the frequency of the movement of the nuclei:

$$\mu(t) = E_0[\alpha_0(\nu) + \alpha'(\nu)\cos 2\,\pi\nu' t]\cos 2\pi\nu' t \tag{7.16}$$

This can be written as:

$$\mu(t) = E_0\alpha_0(\nu)\cos 2\pi\nu t + E_0\alpha'(\nu)\cos 2\pi\nu' t \cdot \cos 2\pi\nu t \tag{7.17}$$

The first term is the normal induced dipole scattering, while the second corresponds to shifted frequencies. One can now use the identity $\cos A \cdot \cos B = ½[\cos(A+B) + \cos(A-B)]$ to obtain:

$$\mu'(t) = E_0\alpha(\nu)[\cos 2\pi(\nu - \nu')\cdot + \cos 2\pi(\nu + \nu')t] \tag{7.18}$$

This equation describes two shifted frequencies at longer wavelength ($\nu - \nu'$) and at shorter wavelength ($\nu + \nu'$) in the scattered light around the exciting wavelength. Although it may look strange that a sample is excited with light and that scattered light contains photons with higher energy, it is more plausible to see the Raman Effect as a two-photon process: the molecule emits the incoming light quantum plus a vibrational quantum.

7.5.2 Raman Experiments

In a Raman experiment, a sample is illuminated with intense monochromatic light and the scattered light is analyzed with respect to its frequency and intensity by a monochromator. If the intensity of the scattered light is plotted with respect to the difference frequency (exciting light-minus-emitted light), the Raman spectrum is obtained. Similarly to the IR spectra discussed above, it is composed of vibrational and rotational transitions.

Upon planning a Raman experiment, it should be considered that the Raman Effect is a scattering process with extremely low probability – only one of 10^{10} photons is reemitted with a shifted frequency. To obtain a measurable number of photons, the intensity of the exciting light needs to be very high and this usually requires lasers.

The intensity of the scattered light shows a maximum at the frequency of the exciting light; this is called the Rayleigh line because of elastic Rayleigh scattering. Typically, it is several orders of magnitude stronger than the Raman lines, depending on the sample state and on aggregates in the sample (which increase Rayleigh scattering).

For the practical use of Raman spectroscopy, most experiments are designed to observe the Stokes lines shifted to lower energy. Because of the occupation of vibrational levels, the anti-Stokes lines are extremely weak. The intensity of the scattered radiation is a function of the excitation wavelength; it varies with λ^{-4}. Consequently, this means that excitation with higher energy photons (blue light instead of red light) will lead to a higher Raman intensity.

Figure 7.28 schematically shows an optical setup for a Raman spectroscopy. Light from an intense light source at a wavelength λ is focused on a sample; the photons emitted are collected, spectrally separated, and detected. Similarly to fluorescence experiments, emitted light is detected at 90° to reduce scattering. A double or triple monochromator is necessary to suppress the high intensity of the Rayleigh line and to scan the wavelengths close to the excitation line.

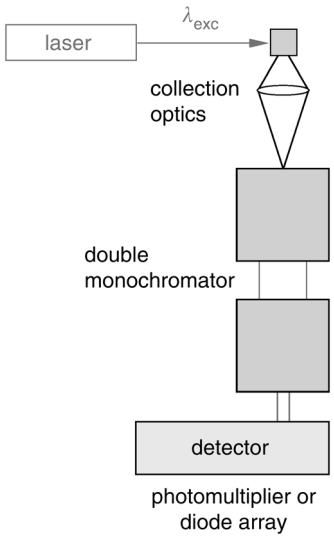


Figure 7.28 Schematic optical setup of a Raman experiment.

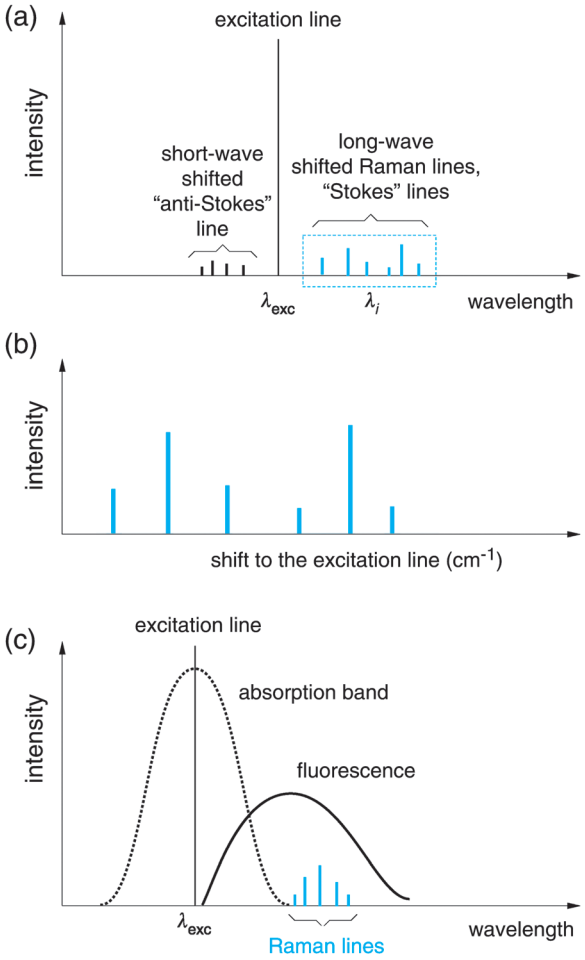


Figure 7.29 Measured spectrum consisting of scattered lines (a), and the Raman spectrum (b) calculated from these. (c) Spectrum showing the overlap of the excitation line with an absorption band and the possible overlap of an emission band with Stokes shifted Raman lines.

Many biological samples show fluorescence upon excitation as soon as the excitation light approaches an absorption band. This can be used for resonance enhancement in resonance Raman spectroscopy (see below). Figure 7.29 shows how fluorescence can overlap the weak Raman lines. It is necessary to keep in mind the orders of magnitude: For some biological chromophores with fluorescence quantum efficiency close to 100%, almost each photon is reemitted. The probabilities for emitting an inelastically scattered photon are of the order of 10^{-8}–10^{-10}.

Raman studies on proteins can use excitation wavelengths in the visible part of the spectrum or in the UV. No absorption is required because the effect relies on scattering. The excitation wavelength is thus frequently selected according to the possibilities for detection; blue light is preferred because of the higher scattering intensity for shorter wavelengths. It should be considered, however, that excitation at high intensities can also cause unwanted photochemical reactions in a biological sample.

7.5.3 Resonance Raman Spectroscopy

Raman spectroscopic studies on proteins without chromophores in the visible or the UV part of the spectrum have always been limited. With the advent of tunable lasers it became possible to tune the excitation light wavelength onto the absorption bands of electronic transitions arising from chromophoric groups. This procedure is called resonance Raman Spectroscopy. The resonance effect drastically increases the polarizability of the molecule, which increases the intensity of the scattered radiation. This can be used to separate the Raman lines of the respective chromophore group from the background of the Raman lines arising from the entire molecule. Resonance Raman scattering has proven to be a valuable tool for pigment–protein systems, for example, there are abundant data on retinal proteins or chlorophyll–protein complexes.

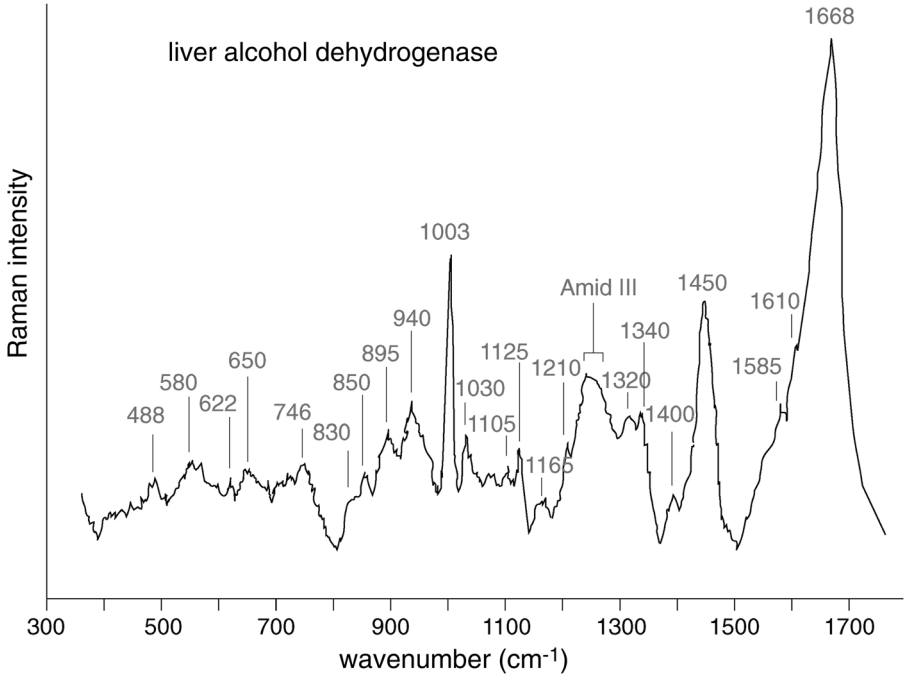


Figure 7.30 Raman spectrum of liver alcohol dehydrogenase.

Some lasers emit in the UV; they can be used for Raman studies with resonant UV excitation. This can be of advantage in the study of proteins devoid of pigments by tuning into the absorption bands of the aromatic amino acid side chains near 280 nm.

Figure 7.30 shows the Raman spectrum of liver alcohol dehydrogenase (LADH). The spectral range 1700–300 cm^{-1} shows numerous Raman bands that can be used in detail for conformational analysis of NADH in the free and in the bound state.

We have discussed above the multiplex advantage of Fourier transform infrared spectroscopy. A parallel method has been established in Raman spectroscopy, where the emitted Raman scattering is not analyzed in a dispersive way, that is, using a monochromator as in Figure 7.28, but by using an interferometer for simultaneous detection of all wavelengths. The multiplex advantage facilitates the detection of extremely weak Raman scattering upon long-wavelength excitation. The combination of elements from FT-IR spectroscopy, that is, the interferometer, and the neodymium-YAG laser emitting at 1064 nm, have led to the name NIR-FT-Raman spectroscopy. This technique is now widely used and benefits from the excitation with low-energy photons that do not cause photoreactions, and are outside the absorbance bands for most biological molecules. However, despite this excitation, the multiplex advantage with interferometric detection yields low-noise spectra.

Raman spectra are analyzed analogously to IR spectra. The position of Raman bands can be interpreted in terms of normal modes or of group vibrations (see above). Assignments are obtained, for example, by introducing isotopes that alter the mass in Equations 7.11 and 7.12 and thus shift the frequency.

For IR spectroscopy, the prerequisite for activity of a molecule is the oscillation of a dipole moment with the vibration. In Raman spectroscopy, the polarizability has to change with the movement of the nuclei. For simple molecules, both are complementary, meaning that a vibrational mode is either IR active or Raman active. For complex biological molecules, this simple symmetry rule is almost always violated and Raman intensities and IR absorbance are observed in parallel.

7.6 Single Molecule Spectroscopy

All spectroscopic methods discussed here probe a large statistical ensemble. If one assumes concentrations of some micromoles per liter, typical sample volumes of a µl still contain around

10^{10} molecules. The properties of the molecules probed with any of the techniques in this chapter all represent a statistical average. However, it is equally relevant to know the properties of single molecules.

Single molecules can be studied by fluorescence techniques, provided they are marked with fluorescent probes that have a high quantum efficiency. Unfortunately, the intrinsic fluorescence of polypeptides is not strong enough.

Single molecule spectroscopy uses, for example, a confocal microscope where the illuminating optics and the microscope objective have a common focus that allows observation of a measuring volume of the order of 10^{-15} liter (1 femtoliter; 1 µm × 1 µm × 1 µm). Provided that the measuring solution is sufficiently dilute, single molecules are detected and, if immobilized, can be observed and analyzed for a longer period. Excitation is performed with a laser at fixed wavelength.

Confocal Laser Scanning Microscopy, Section 8.6

Fluorescence measurements for a large ensemble exhibit constant fluorescence over longer periods, provided that the fluorescence label does not bleach and the absence of reactions or diffusion make for a constant concentration. In contrast, single molecule fluorescence does not yield continuous emission, but rather a sequence of *on*/*off* states of individual dye molecules. This blinking, including longer *off* times, has different causes: A dye molecule can "park" with a certain probability in the triplet system (Figure 7.6), which leads to dark intervals of up to 100 ms. Blinking itself is taken as a proof that single molecules are in the measuring volume. A second cause of the stopping of fluorescence is photobleaching.

Fluorescence signals are evaluated with statistical methods, for example, by measuring the occurrence of the *on*-states as a function of their duration. Typically, an exponential decay is observed, which is taken as a hint for a Poisson distribution of the *on*-state.

Today, many cell biology techniques are based on the measurement of concentration and diffusion of fluorescent molecules by single-molecule techniques. In fluorescence correlation spectroscopy (FCS, see Section 7.3.6) the fluctuations that occur when fluorescent particles diffuse through the measuring volume are analyzed. By using a CCD camera, instead of the diffusion of local intensity traces, the trajectories of fluorescent labeled particles are measured. This is called single particle tracking and can be used to differentiate between normal diffusion and directed transport, for example, due to a gradient. Single particle tracking is a valuable tool for the investigation of dynamic processes in living cells.

Single molecule spectroscopy is highly attractive for the analysis of oligomeric aggregates. As an example we take the photosynthetic light-harvesting complexes where apparently identical chlorophyll–protein complexes surround the reaction centers (absorption spectra in Figure 7.19). In photosynthetic purple bacteria, light-harvesting complex B870 (absorption and emission spectra shown in Figure 7.19) consists of a ring of 32 bacteriochlorophyll *a* molecules. Single molecule spectroscopy has unraveled the properties of the individual complexes and has helped to clarify the processes of energy conduction within this ring and from the ring to the reaction center inside.

7.7 Methods using Polarized Light

The methods discussed up to now have made use of unpolarized light. If we add polarization as a parameter, we can get additional information on the conformation of biological macromolecules and on their orientation.

Essentially, three methods using light polarization find applications in bioanalysis: linear dichroism, optical rotation dispersion, and circular dichroism. The second and third techniques apply for the investigation of optically active molecules.

7.7.1 Linear Dichroism

Linear dichroism makes use of the orientation of the transition dipole moment (Section 7.1) within the geometry of the molecule. As discussed in Section 7.1, the electric field vector of the incoming wave interacts with this transition dipole moment. The strength of interaction depends on the relative orientation of the field vector and the transition dipole moment: if parallel there is maximum interaction, if perpendicular it is zero. The contribution of a molecule to absorption

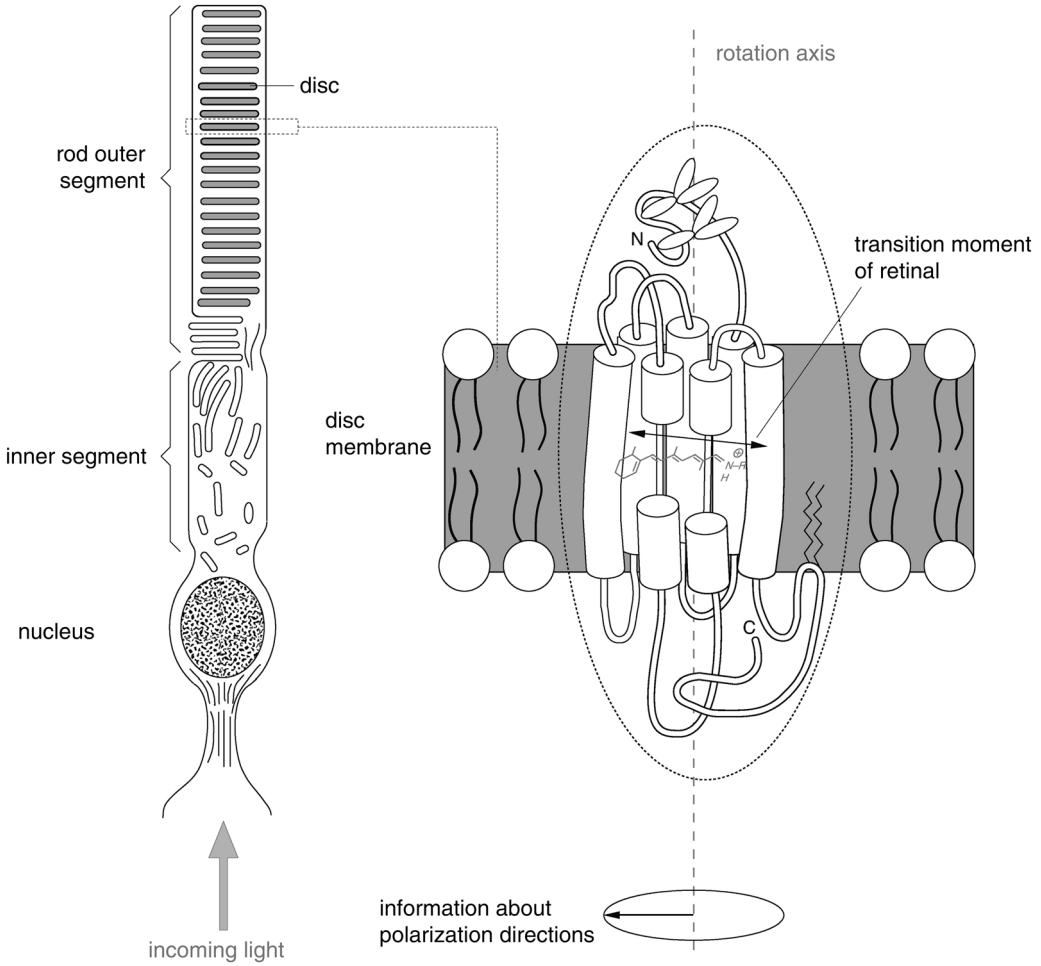


Figure 7.31 Arrangement of the "absorber" retinal in the visual pigment rhodopsin relative to the incoming light.

thus also depends on the orientation of the absorber with respect to the field vector. Light as a transverse electromagnetic wave consequently cannot monitor molecules with the transition dipole moment oriented in the direction of propagation.

In a large ensemble of molecules that can diffuse and rotate, this dependency on orientation perfectly averages out (remember that there are around 10^{10} molecules present), and the same absorption is measured for any direction of the electric field vector. This is called an isotropic sample. Biological samples, however, show oriented chromophores that exhibit absorption values that may strongly depend on the polarization of the light. Linear dichroism spectroscopy uses this effect to probe the orientation of chromophoric groups or pigments in macromolecules or in biomembranes.

The human eye is the classical case of an absorption process optimized for all directions of polarization of the incoming light (Figure 7.31). In the retina, the rod cells are oriented to have the incoming light exactly along their axis. Light thus enters perpendicular to the approx. 2000 discs in the rod outer segment that form a stack like CDs on a rack. These discs are flattened vesicles that contain in their membrane the visual pigment rhodopsin with retinal as the chromophore. Rhodopsin is mobile in the disc membrane and can diffuse laterally as well as rotate. However, its long axis essentially remains parallel to the direction of arrival of the light. The absorbing chromophore retinal is approximately perpendicular to this long axis, and thus is with its transition moment in the plane of the electric field vector. Rotational diffusion of rhodopsin allows all orientations in this plane and thus ensures efficient absorption for all polarization directions.

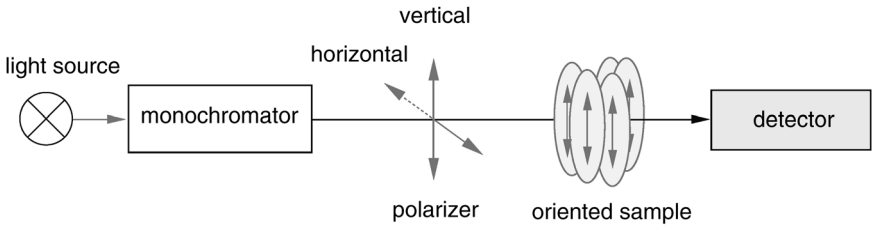


Figure 7.32 Setup for a linear dichroism experiment.

A linear dichroism experiment is rather simple and requires a polarizer and an oriented sample in addition to a spectrometer. Typically, two spectra are recorded (Figure 7.32), one with vertical and one with horizontal polarization. The measured quantities are $I_{\mathrm{vert}}(\lambda)$ and $I_{\mathrm{hor}}(\lambda)$. The reference spectra with a suitable reference sample also need to be recorded for both polarization directions: $I_{0,\ \mathrm{vert}}(\lambda)$ and $I_{0,\ \mathrm{hor}}(\lambda)$, because almost all components of a spectrophotometer exhibit some "prepolarization" and thus different transmission for both polarization directions. After that, absorption can be calculated:

$$\begin{aligned} A_{\mathrm{vert}} &= \log\left(\frac{I_{0,\mathrm{vert}}}{I_{\mathrm{vert}}}\right) \\ A_{\mathrm{hor}} &= \log\left(\frac{I_{0,\mathrm{hor}}}{I_{\mathrm{hor}}}\right) \end{aligned} \tag{7.19}$$

The difference between both absorbance values is called the linear dichroism (LD) of a sample; it yields information on the orientation of a pigment in a biological structure in addition to its absorption.

Biological samples for LD experiments can be orientated by different methods. Lipid membranes can be oriented by careful drying on a glass or quartz support with their plane parallel to this, but are upside–down statistically distributed. This *uniaxial* orientation samples are like coins dumped on a table. Other techniques for orientation imply electric or magnetic fields, or orientation by hydrodynamic forces in a flow chamber. The ideal orientation is in crystals. LD microspectroscopy on small protein crystals can be used to study crystal quality and space group if these crystals are still too small for X-ray analysis.

The relative orientation of a transition dipole vector and the electric field not only determines the absorption strength of electronic transitions but also of vibrational transitions. In a C═O bond, the transition dipole moment is oriented along the bond axis, and even for more complicated geometries the IR transition dipole moment is fixed in the molecule geometry. This offers a simple method to study the orientation of protein secondary structure elements (Figure 7.33).

In an α helix, there are hydrogen bonds from the peptide N─H bond to the C═O group (Figure 7.33) that form an almost linear ─C═O · · · H─N─ structure approximately aligned with the axis of the helix. Similar H-bonding is formed in β-sheet structures, but at different H bond strength and thus with different ─C═O frequency; in this case, the ─C═O · · · H─N─ structure is approximately aligned perpendicular to the sheet axis. We have seen in Section 7.4.4 how the infrared absorption from different types of H-bonded C═O vibrations from the amide I mode can be used for the analysis of secondary structure elements and for folding/unfolding processes. Beyond that, the measurement with polarized IR light can be used to determine the orientation of these secondary structure elements.

A sample for this orientation analysis can be a membrane sheet that has been oriented by gentle drying on an IR transparent support (e.g., a CaF_2 disc). For proteins embedded in this membrane, the transmembrane α helices spanning the membrane should exhibit maximum absorbance if the electric field vector is aligned along the membrane normal (which is an axis perpendicular to the supporting disc). This implies inclination of the sample with respect to the beam. The amide I absorption with the IR beam polarized horizontally and vertically can then be used to calculate the maximum inclination of the helix axis with respect to the membrane normal – an important piece of information in the primary characterization of membrane proteins.

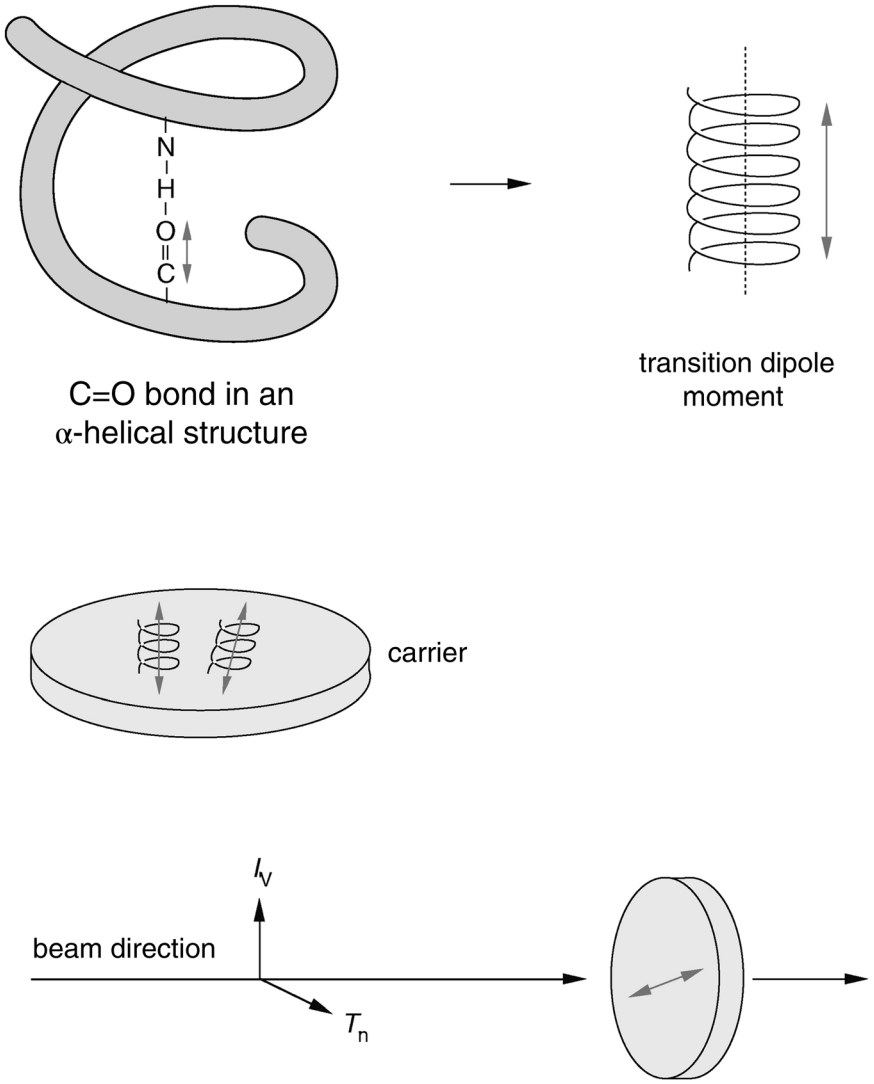


Figure 7.33 Linear dichroism IR spectroscopy for the determination of α helix orientation in membrane proteins.

7.7.2 Optical Rotation Dispersion and Circular Dichroism

Optical rotation dispersion (ORD) and circular dichroism (CD) use the interaction of optically active molecules with polarized light – both methods monitor the same phenomenon, but from different perspectives. ORD measures the difference in refractive index of an optically active compound for left and right circular polarized light, while CD measures the difference in absorption.

Optical activity arises with the introduction of a chiral center in a molecule, for example, if a carbon atom binds four substituents in a tetrahedron form. In an optically active compound like this the speed of light and thus the refractive index is different for left and right circular polarized light: n_L and n_R. One light component is thus delayed after passing through an optically active sample. Because linear polarized light can be described as composed of two superimposed left and right circular polarized components, the delay after passing the sample can be registered as a rotation of the polarization plane of linear polarized light upon passing through a sample.

This rotation is proportional to the path length and to the concentration of the optically active substance. A well-known application is the measurement of sugar concentrations with a polarimeter ("saccharimeter"). The rotation of the polarization plane is usually referred to molar concentration and to path lengths of 1 cm.

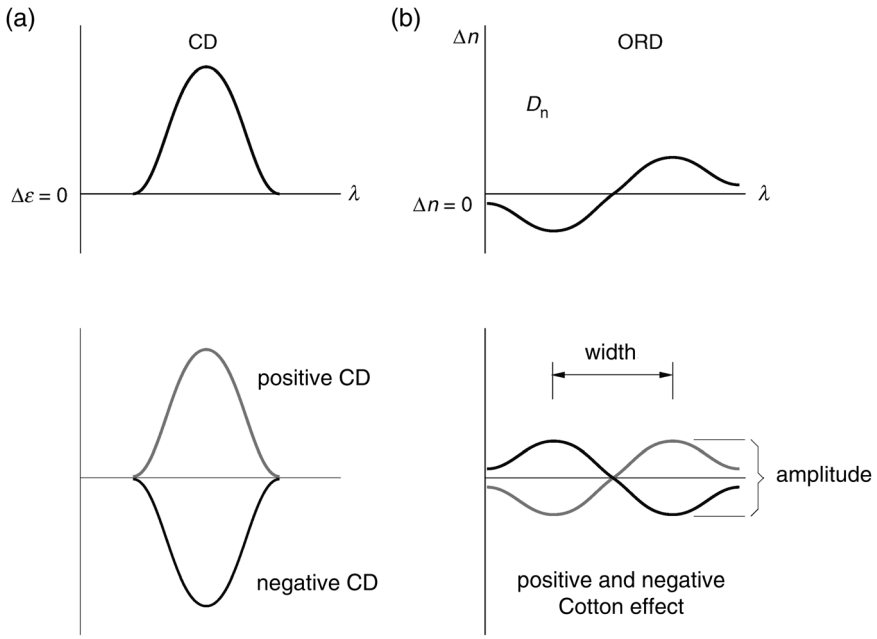


Figure 7.34 Circular dichroism (a) and optical rotation dispersion spectrum (b). For CD, the difference $\Delta\varepsilon = \varepsilon_L - \varepsilon_r$ is displayed, for the ORD spectrum the difference is between n_L and n_R. Note that the CD spectrum in (a) would correspond to a normal absorption spectrum if the molecule did not exhibit chirality (in which case, $\Delta\varepsilon = 0$). For (b), a non-chiral molecule would yield a "normal" dispersion curve ($\Delta n = 0$).

Circular dichroism (CD) spectroscopy also relies on the overlap of a left and a right circularly polarized light wave. In contrast to ORD, where n_L and n_R are analyzed, it measures the difference in absorption of the left and right circularly polarized components of light. In Section 7.2 we introduced the molar absorption coefficient ε for the characterization of the absorption properties of a molecule; for CD we extend this to ε_L and ε_R for the absorption properties of the molecule for left and right circularly polarized light, respectively. The difference $\Delta\varepsilon = \varepsilon_L - \varepsilon_R$ is measured and given as the *ellipticity* Θ:

$$\Theta(\lambda) = \text{const}(\varepsilon_L - \varepsilon_R)cd \tag{7.20}$$

where d is the path length and c is the concentration of the sample. The ellipticity $\Theta(\lambda)$ is recorded in a CD-spectrum.

Figure 7.34 shows a schematic CD spectrum (a) and an ORD spectrum (b). The CD spectrum shows the difference in the molar absorption coefficients versus the wavelength; in the ORD spectrum, the difference in refractive index is shown. The spectra arise from an absorption band in an optically active chromophore. Spectra in this form are also known as the Cotton effect. They can be positive or negative as shown in Figure 7.34. The two enantiomers of a molecule exhibit the same Cotton effect, but with opposite sign.

The differences in refractive index for ORD spectroscopy and in absorption for CD spectroscopy are small with respect to the total refractive index or the total absorption. ORD spectroscopy of amino acids yields rotation angles of only a few degrees for 1 mM solutions and path lengths of a few cm. The difference in absorption is typically 10^{-3} of the total absorption or even less. Technologies for the measurement of CD spectra are thus more elaborate. First, a monochromator is used to select light of a wavelength λ that is then linearly polarized. To generate an alternating left- and right circularly polarized wave, a photoelastic modulator is used at high frequencies. The detector is synchronized with this modulator and detects alternating I_L and I_R as the basis for calculation of the ellipticity Θ or, upon variation of the wavelength, a CD spectrum.

The most relevant application of CD spectroscopy is the analysis of protein secondary structures. This method is based on the UV absorption of the peptide bond in the spectral region 160–250 nm that we have already discussed in Section 7.2.1. The absorptions of the peptide bond appearing in this range are the $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ electronic transitions, which are weak in the absorption spectrum and not really characteristic for the secondary structure. As absorption bands, they are not used for analytical purposes. However, the CD spectrum in this part of the

spectrum is extremely sensitive to the secondary structure because of the chirality of these structures. This can be shown for model peptides that exhibit different secondary structures that are dependent on pH and temperature. Poly-L-lysine, for example, exhibits an α-helical structure at pH values below 10 and an unordered structure above this pH; this transition is perfectly reflected in the CD spectrum.

To analyze the secondary structure of an unknown by UV-CD spectroscopy, a CD spectrum is first taken in a buffer that does not exhibit strong background absorption. It is also necessary to verify that other components that come with the protein, such as, for example, detergent or urea (in the case of unfolding studies), do not absorb strongly. In a second step, a linear combination of CD spectra from (ideally) pure α-helical, β-sheet, and random coil structures is used to model the measured CD spectrum. The fraction percentages of the "pure" spectra are then the fractions of the unknown protein. A more elaborate data evaluation uses a basis set of CD spectra from water soluble and membrane proteins with different secondary structures to model the unknown protein. Some of these data bases are accessible on the internet (http://dichroweb.cryst.bbk.ac.uk), where the measured CD spectrum can be uploaded and the secondary structure analysis is obtained. Examples for CD spectra of proteins and peptides are given in Chapter 22.

This secondary structure analysis works quite well, but has some pitfalls, as already seen for the IR-based secondary structure analysis. These are mostly a strong background absorbance by urea, detergent, or buffers and a lack of matching between the wavelengths ranges for the measured protein and the ones used as basis set. However, whenever this analysis is carefully performed and eventually combined with IR spectroscopy, a consistent picture of the secondary structure of unknown proteins can be obtained.

Cotton effects from nucleic acids appear in the range 250–275 nm. They arise from electronic transitions of the nucleotide bases. CD spectroscopy can also be used for the visible spectral range in order to study the coupling of chromophores. This close coupling of chromophores leads to either a lowering or an increase of electronic levels and thus to clear positive/negative bands in the CD spectrum.

Further Reading

Cantor, C.R. and Schimmel, P.R. (1980) *Biophysical Chemistry Part II: Techniques for the Study of Biological Structure and Function*, W.H. Freeman and Company, New York.

Cid, M. and Bravo, J. (eds) (2015) *Structure Elucidation in Organic Chemistry: The Search for the Right Tools*, Wiley-VCH Verlag GmbH, Weinheim.

Colthup, N.B., Daly, L.H., and Wiberley, S.E. (1990) *Introduction to Infrared and Raman Spectroscopy*, Academic Press.

Diem, M. (2015) *Modern Vibrational Spectroscopy and Micro-Spectroscopy: Theory, Instrumentation and Biomedical Applications*, John Wiley & Sons Ltd, Chichester.

Mäntele, W. and Deniz, E. (2017) UV–VIS absorption spectroscopy: Lambert-Beer reloaded. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Vol. 173, 965–968.

Mäntele, W. (2015) The analysis of protein conformation by infrared spectroscopy: A Tutorial. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Vol. 138, 964–966.

Mäntele, W. (2012) The seven sins of fluorescence measurements. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Vol. 96, V–VII.

Schmidt, W. (2005) *Optical Spectroscopy in Chemistry and Life Sciences*, Wiley VCH Verlag GmbH, Weinheim.

Siebert, F. and Hildebrandt, P. (2008) *Vibrational Spectroscopy in Life Science*, Wiley VCH Verlag GmbH, Weinheim.

Turro, N.J. (1991) *Modern Molecular Photochemistry*, University Science Books, Mill Valley, CA.

Light Microscopy Techniques – Imaging

8

Thomas Quast,[1] Mike Heilemann,[2] and Waldemar Kolanus[1]

[1]*University of Bonn, Life and Medical Sciences Institute (LIMES), Carl-Troll-Straße 31, 53115 Bonn, Germany*
[2]*Johann Wolfgang Goethe University, Institute of Physical and Theoretical Chemistry, Max-von-Laue Straße 7, 60348 Frankfurt, Germany*

8.1 Steps on the Road to Microscopy – from Simple Lenses to High Resolution Microscopes

In his main work the *Book of Optics* the polymath Alhazen (964 to ca. 1040) described the phenomena of light and reflection as well as the characteristics of spherical and parabolic mirrors. In medieval times, Roger Bacon (1214–1292) developed optical instruments and lenses to correct vision. The beginnings of light microscopy go back to the seventeenth century. Inspired by the great success of the telescope, which Galileo Galilei (1564–1642) used in astronomy, experiments were begun to magnify small nearby objects to the greatest possible extent.

Johannes Kepler (1571–1630) discovered the laws of total reflection and refraction for small angles of incidence. He also built the first microscope made of two convex lenses. Robert Hooke (1635–1703), Antoni van Leeuwenhoek (1632–1723), and Marcello Malpighi (1628–1694) are the outstanding figures who paved the way for actual light microscopy.

Hooke developed a microscope consisting of two lenses, which he used to investigate fine slices of corkwood, among other things. He discovered tiny chambers in the wood that he described as cells in his main work *Micrographi*a (Figure 8.1). To illuminate the preparation he used the light of an oil lamp surrounded by a glass globe filled with water, which served to distribute the light of the lamp evenly. He constructed lens grinders, which, however, did not lead to the desired resolution of the fine details. In addition, Hooke's microscope had the disadvantage that distortions produced by spherical aberrations were amplified by the combination of lenses. Much later it was discovered that the spherical aberrations that lead to the distortions can be reduced by the use of filters. Chromatic aberrations cause colored edges of preparations and only much later was is possible to correct these by the use of special combinations of lenses.

As a result, until the nineteenth century, simple microscopy, employing only a single lens, was standard practice. Small lens diameters allow very high curvature and therefore very short focal points. Leeuwenhoek used tiny, bubble free single lenses (Figure 8.2), which he painstakingly ground down until he was able to achieve an almost 270 fold magnification. In addition to investigating blood and capillaries of tadpoles, he discovered protozoa. Drawings of objects moving under their own power, sent in a letter to the Royal Society of London in 1675, suggest that his was the first eye in history to see what would later become known as bacteria (Figure 8.3).

Malpighi's investigations focused primarily on plants and insects. During these investigations he discovered the lung of the frog in a maze of blood vessels.

At the end of the nineteenth century, Ernst Abbe succeeded in significantly improving the design of microscopes with the aid of investigations into the effects of the bending of light waves (diffraction) and the calculation of the level of magnification. The company Carl Zeiss

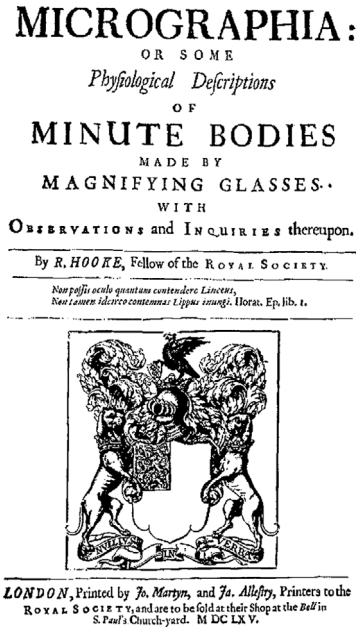

MICROGRAPHIA:
OR SOME
Physiological Descriptions
OF
MINUTE BODIES
MADE BY
MAGNIFYING GLASSES.
WITH
OBSERVATIONS and INQUIRIES thereupon.
By *R. HOOKE*, Fellow of the ROYAL SOCIETY.
Non poſſis oculo quantum contendere Lincei,
Non tamen idcirco contemnas lippus inungi. Horat. Ep. lib. 1.
LONDON, Printed by *Jo. Martyn*, and *Ja. Allestry*, Printers to the ROYAL SOCIETY, and are to be ſold at their Shop at the *Bell* in S. *Paul's* Church-yard. M DC LX V.

Figure 8.1 Title page of the major work "Micrographia: or some physiological descriptions of minute bodies" Robert Hooke, 1565. SLUB Dresden, CC-BY-SA-4.0.

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

Figure 8.2 Microscope constructed by the Dutch merchant Antoni van Leeuwenhoek (1632–1723). The microscope consists of a tiny lens or magnifying glass, as used by cloth makers at that time to determine the quality of the material. The lenses, which Leeuwenhoek ground himself, were clamped into metal frames. With the aid of this simple microscope Leeuwenhoek was able to observe objects at a magnification of 270×, which was far beyond the capabilities of the first multiple lens microscopes. Source: Deutsches Museum, München, Germany.

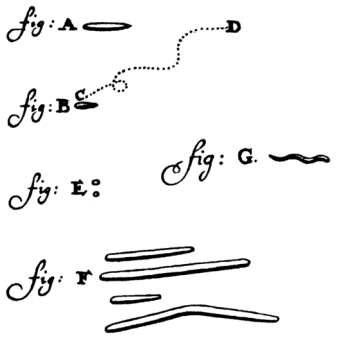


Figure 8.3 First picture of bacteria. During the investigation of dental plaque, Antoni van Leeuwenhoek was apparently the first person to discover bacteria. He sent the drawing in the figure showing bacteria moving under their own power in a letter to the Royal Society of London at the end of the sixteenth century. Source: The Royal Society, CC-BY-4.0.

Electron microscopy, Chapter 19

manufactured a series of objectives, beginning in 1886, according to Abbe's calculations, which enabled anatomists to recognize objects close to the theoretical limit of resolution of light microscopes. The maximum resolution of a visible light microscope is limited by the wavelength of the light used, which approximately spans the range from 400 nm (violet) to 700 nm (dark red).

Over the course of time, the microscope became an important scientific tool. A milestone in this development was the theory that plants and animals consist of cells, advanced by Matthias J. Schleiden and Theodor Schwann in 1838/39. Cell biological research began with light microscopic techniques, which have been developed, along with electron microscopic procedures, to essential, flexible visualization methods for bioanalysis.

With the help of aniline dyes, Robert Koch (1843–1910) was able to stain microorganisms and identify the causes of tuberculosis and cholera. With the inventions of the phase contrast microscope by the Nobel Prize laureate Frits Zernike (1888–1966), interference microscopy by Lebedeff, and differential interference contrast microscopy developed by Georges Nomarski (1919–1997), microscopy of unstained, living cells became possible.

The specific labeling of cellular antigens with fluorescent antibodies was first accomplished by Albert Coons in 1941. The principle of confocal microscopy was patented in 1957 by Marvin Minsky. Using these fundamental principles, David Egger and Mojmir Petran built the first confocal laser scanning microscope (LSM) on the basis of a spinning disk (Nipkow) system (Section 8.6). As a result of the combination of improved computer and laser technology with new algorithms for digital processing of pictures, the commercial applications of LSM have greatly expanded since the mid-1980s. The newest developments are focused primarily on the improvement of resolution and the acceleration of image acquisition. With the aid of controlled environment incubators (control of temperature, CO_2, humidity), motorized microscope control, and high resolution, sensitive detection systems (digital cameras, photomultipliers), it has become possible to visualize tissues as well as cells and the dynamics of their molecular components live over extended periods of time. Modern digital photo-processing programs allow 4D reconstructions and the quantitative evaluation of time courses.

The possibilities of fluorescence microscopy were significantly improved with the discovery of green fluorescent protein (GFP), for which Osamu Shimonmura, Martin Chalfie and Roger Tsien received the Nobel Prize for Chemistry in 2008. GFP was expressed for the first time outside of the fluorescent jellyfish *Aequorea victoria* (Figure 8.7a below) in 1994, which was a significant step on the road towards using the protein as a genetic marker. The extremely stable GFP (Figure 8.7b,c) can be fused with any protein and, after expression, allows the specific spatial and temporal localization of proteins in living cells. In the meantime the gene has been extensively modified to obtain proteins that fluoresce in other colors. In this manner, great flexibility in the simultaneous localization of fusion proteins has become possible (Section 8.4).

8.2 Modern Applications

The great variety of available techniques and the capabilities of high resolution visualization and subsequent computer-aided image processing make light microscopes useful for material sciences, as well as life sciences. Other uses include medical diagnostics, quality control in the pharmaceutical and food industries, and in the areas of soil and water sciences. The light microscope is no longer limited to just structural localization and characterization of certain subcellular components in fixed cells with the aid of special detection methods (e.g., fluorescent labels, chemical stains), but instead can involve the investigation of dynamic processes in living cells and tissues. This was made possible primarily by the automation of image acquisition and the computer-aided motorization of the microscope controls such as stages, objective, autofocus, fluorescence filter, and light shutters. This technique, referred to as live cell imaging, makes it possible to detect very slow and very rapid processes in living cells in real time (Section 8.6).

Electron microscopic techniques enable a significantly higher resolution but cannot be used to observe living cells, thus reducing their usefulness. In addition, the preparation techniques for electron microscopy are usually much more laborious than for light microscopy.

Furthermore, special developments in the area of live cell imaging (FRET, FLIP, FRAP, single molecule microscopy; Section 8.6) allow the investigation of fluorescently marked molecules with the aid of lasers and the analysis of interactions at the molecular level in real time (Section 8.6).

Fluorescence Spectroscopy, Section 7.3

Light microscopic technologies are also very important in medical diagnostics, in particular in the pathological analysis of tissue biopsies (e.g. chemical staining of paraffin-embedded slices) and blood. Neuro- and microsurgical operations make use of high resolution optical systems. An interesting application of *in vivo* confocal LSM is, for example, the investigation of basal cell carcinomas. With a maximum depth of penetration of 200 µm in human skin, this procedure offers pre-operative diagnostic possibilities in which changes in the tissue or the vasculature can be analyzed to determine if the area has become a carcinoma. These procedures are already in use in dentistry for the inspection of implants in order to analyze the contact zone between tissue and dental implant.

8.3 Basic Physical Principles

To better understand the physical basis of a microscope and the necessary steps (preparation and analysis), a few important concepts will be explained in the following glossary.

Microscopic Glossary

Achromatic objectives
Achromatics are the most common objectives in microscopy. The imaging errors of these objectives are not corrected to the same degree as the plan achromatics and the plan apochromatics. Owing to their simpler mechanisms and optical components, achromatics are always among the least expensive objectives of a manufacturer and they are well suited to many routine microscopic investigations.

Apochromatic objectives
Apochromatic objectives use a complicated arrangement of special lenses in the objective to suppress the colored edges of a microscopic image.

Aperture diaphragm
The aperture diaphragm is an iris diaphragm below the condenser. With this diaphragm, a compromise is achieved between resolution and contrast (so-called illumination aperture). Opening the diaphragm increases the resolution while decreasing the contrast; correspondingly, closing the diaphragm leads to increased contrast, but often at the price of lower resolution. Since the brightness of the microscopic image changes at the same time, the aperture diaphragm is frequently misused to regulate brightness. This is a common mistake in the use of microscopes. When done properly, the brightness is reduced either by a regulator or through filters.

Free working distance
The distance between the front lens of the objective and the surface of the carrier of the preparation (e.g., cover slip, cell culture dish), when the preparation is sharply focused, is referred to as the free working distance. Strongly magnifying objectives with high resolution have the shortest free working distances, sometimes below 0.20 mm. To protect the objective and the sample, such objectives usually have a spring-mounted front lens. Therefore, these objectives cannot be used in situations that require a larger free working distance, such as for the microscopic examination of cells on the bottoms of plastic dishes.

Köhler illumination optimization

A method published by August Köhler (1866–1948) arranges the illumination of a microscope in such a manner that, under maximal lighting conditions, a homogenous lighting of a preparation is achieved.

Binocular tubes

Binocular tubes enable the observation of a preparation with both eyes at the same time. This allows no more to be seen as with a monocular tube, nor does it confer depth of field. However, microscopic investigations with such tubes are less tiring and more comfortable than using simple monocular tubes.

Cover slip

In general, a microscopic preparation is transferred to a glass slide and subsequently covered with a cover slip. Particularly with strongly magnifying objectives, the preparation has to be covered for optical reasons (see Section 8.5). The cover slip has a different refractive index than air and therefore influences the path of the light coming from the preparation. This phenomena must be taken into account in the construction of an objective. Usually, objectives are corrected for cover slips with a thickness of 170 µm.

Interference

Interference describes the overlapping of two or more waves coming from a preparation, which leads to the addition of their amplitudes (not their intensities). Constructive interference refers to amplitude amplification, while destructive interference refers to waves that cancel each other out.

Microscopic stages

For the systematic observation of a preparation with limited vibrations, a microscopic stage, also referred to as an X,Y table, is necessary to hold the preparation. This allows a very precise manipulation of the sample in the x, y, and z directions. Motorized stages are now commonly available that allow an automatic position change in three dimensions.

Condenser

In inverted microscopes, the lens system is below the sample, on the top of which is an adjustable iris diaphragm (aperture diaphragm: only in bright field condensers – dark field condensers lack an aperture diaphragm). These components together comprise the condenser. The microscope light first goes through this condenser before it strikes the sample. The function of the condenser is the optimal preparation and regulation of the light entering the microscope. The illumination system of a microscope therefore consists of a light source (e.g., halogen lamp) and the condenser. The light source is responsible for the necessary brightness (light quantity) and the condenser for the necessary concentration of light (light quality).

Coherent light

The designation coherent comes from the optical wave model of light. Light waves, which fulfill the following criteria, are described as coherent:

- same wavelength
- same polarization
- strike the same point at the same time.

Only light waves that fulfill the criteria of coherence can interfere with one another.

Light field diaphragm

The light field diaphragm is an adjustable iris diaphragm and is found in microscopes with Köhler illumination in the microscope stand. When the diaphragm is correctly adjusted,

only the section of the preparation currently under observation is illuminated. This protects the preparation from excess exposure to light and minimizes the generation of light diffusion, which can lower contrast.

Parfocal distance
While working with microscopes, frequent changes of objectives of different magnifications are necessary. Modern objectives are designed such that, on switching objectives, the preparation remains sharply focused and does not require extensive correction. Objectives that have this property are referred to as balanced or height corrected.

Phase contrast condenser
A special condenser is required for phase contrast investigations. This type of condenser has a phase diaphragm that replaces the aperture diaphragm of a light field microscope for experiments in phase contrast.

Phase contrast objective
Special objectives are necessary for phase contrast measurements. These objectives have a dark ring immediately behind the focus plane, which is known as a phase ring. The greater magnification of the phase contrast objectives delivers unsatisfactory results when used for bright field microscopy.

Plan achromatic objectives
The plan achromatic lenses are corrected such that the arching of focus across a lens, which lead to minor distortions of a microscopic picture, are completely eliminated.

Plan apochromatic objectives
The plan apochromatic objectives are the most laboriously constructed of all the lens types. For these objectives the arching effects, as for plan achromatic objectives, are largely eliminated. In addition, the red or blue edges are prevented from appearing.

Point spread function (PSF)
The point spread function (PSF) describes the effects of bandwidth limiting factors (diffraction, imaging errors, and others). The function describes how an idealized, point-sized object is imaged by a system.

The closer an object is to the eye, the more details can be distinguished. This principle applies down to a certain threshold, under which the eye can no longer distinctly recognize objects. For an adult, this minimum distance is usually around 250 mm. The unit of measurement for the ability to distinguish two points from one another is known as the resolution. With the aid of magnifying lens, this limit of resolution can be increased. The magnification of a lens can be calculated with the following formula:

$$M_o = \frac{250\,(\text{mm})}{f\,(\text{mm})} \tag{8.1}$$

where M_o is the magnification of the objective, 250 mm is the distance between the eye and the image, and f is the focus length of the lens.

A strong magnification can be achieved by the use of two lenses employed sequentially. A simple microscope is of this arrangement, since it has two lenses (objective and ocular, Figure 8.4a). The objective enlarges the object (O) and forms an enlarged real image (O′) at the focus length of the second lens, in this case the ocular. The ocular then creates an enlarged virtual image (O″) that can be viewed as if it were an image at a distance of 250 mm. As a result,

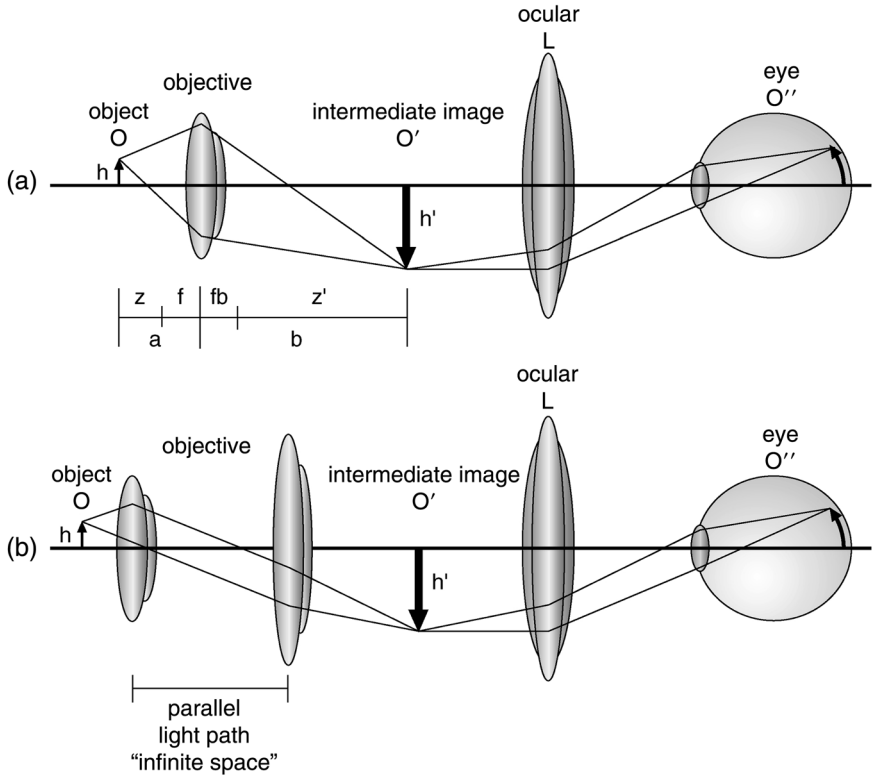


Figure 8.4 Light path of a microscope (a); principle of infinite length optics (b). Modified after Davidson and Abramowitz.

the total magnification of a microscope can be calculated as follows:

$$MA_{Microscope} = M_o M_e \tag{8.2}$$

Infinity corrected optical systems have increasingly become standard among all major manufacturers of microscopes. The light paths in the microscope are altered in these devices such that light beams, after they have left the objective, run in parallel in infinite space. In this infinite space, changes can be made that would require additional tubes when using conventional, finite optics (differential interference contrast, fluorescence microscopy).

The infinite space ends at the lens built into the tube. This lens has the additional function of optimizing the image with respect to its chromatic properties. The construction is also more stable and more flexible since there is no need for the conventional inner tube. The high resolution is, ultimately, limited by the properties of the objective. These characteristics are not only determined by the degree of magnification, but also by the so-called numeric aperture (NA) of the objective. This NA is defined as the product the sine of half the angle α of the maximum cone of light that can be imaged by the objective and the diffraction of the medium with which the object is surrounded (n):

$$NA = n \sin \alpha \tag{8.3}$$

where NA is the numeric aperture, n is the diffraction index of the medium, and sin α is half the angle of the emitted light captured by the objective.

Normally, the diffraction index is $n = 1$ when the object is surrounded by air. Since the angle can never be greater than 90°, the numeric aperture can never be greater than NA = 1 for an objective operated in air. The real maximum value is around NA = 0.95, since the distance between the objective and the surface of the cover slip can never have an actual value of zero. A NA of 0.95 corresponds to an angle α approaching 72°. The numeric aperture can be increased by the use of a medium whose index of refraction is greater than that of air. The use of a special immersion oil with an index of diffraction of $n = 1.515$ has proven to be useful, increasing the NA to about 1.4. Even greater indices of diffraction are not particularly useful, since the

numeric aperture is also limited by the index of diffraction of the objective $n = 1.525$. The resolution d is determined by the wavelength of the light λ and the numeric aperture of the objective ($NA_{Objective}$):

$$d = \frac{\lambda}{2NA_{Objective}} \tag{8.4}$$

where d is the resolution, λ is the wavelength of light used, and $NA_{Objective}$ is the numeric aperture of the objective.

Using light of a wavelength of $\lambda = 550$ nm and an oil-immersion objective with an NA of 1.40 results in the following calculation:

$$d = \frac{550\,(\text{in nm})}{2 \times 1.40} = 200\,\text{nm}$$

Therefore, 200 nm is the highest theoretical resolution that a light microscope can ever achieve, which is in essence the result of the work by Ernst Abbe. The resolution of a light microscope can thus be roughly estimated as around half the wavelength of the light employed. If the limit of resolution of the light microscope is known, the maximum possible magnification can be calculated. A magnification can be viewed as useful or worthwhile when two points that can barely be distinguished from one another can be magnified to the point where the human eye can recognize them as distinct from one another. With distance between the object and the eye of 250 mm, the resolution of the human eye is between 150 and 200 µm. As a rule of thumb, for a useful magnification, the following formula applies:

$$500 \text{ to } 1000\,NA_{Objective} \tag{8.5}$$

where $NA_{Objective}$ is the numeric aperture of the objective.

An objective with $NA_{Objective} = 1.4$ accordingly has a useful magnification of almost 1400-fold.

Diffraction Phenomena and Imaging The phenomena that allows light to be described as both a particle and a wave will not be addressed here in detail. The focus is on the fundamental characteristics of the wave-like nature of light, in order to understand how microscopic images come about. In the process, the function of modern microscopic techniques, such as phase contrast, polarization, and interference microscopy, will be introduced.

Wavelength and Interference of Light The wavelength of visible light lies in the range 400–800 nm. Interference refers to the interaction of light waves with one another that leads to an increase or decrease of the amplitudes of the waves at the resulting inflection points. The extreme case is when two waves mutually cancel one another out (such as in interference microscopy). Diffraction is the partial redirection of a beam of light on the edge of opaque objects.

Light Field Microscopy Light field microscopy belongs to the classical microscopic techniques with which samples can be magnified and imaged as true to reality as possible. So-called amplitude samples, such as stained chromosome spreads or histologic samples, can be studied easily with a conventional light field microscope. For transparent samples, however, a light field microscope has its limits. Such samples can only be examined with a light field microscope with extremely little contrast. As a result, particularly in the twentieth century, many new optical contrast techniques, for example, phase contrast, were developed.

Phase Contrast Microscopy A fundamental problem with the microscopic examination of some biological objects is their low contrast. Only when sufficient contrast exists, or it can be achieved with the use of contrast amplifying stains, can biological structures be imaged with light microscopes. Light absorbing areas of a sample weaken the amplitude of the light waves penetrating through them. These decreases in amplitude are perceived by the human eye as differences in brightness. The transparent areas of the sample allow the light to pass through unhindered. Depending on the characteristics of the sample, the light waves may undergo a change in phase, since their speed through the sample is slowed. It is difficult to detect such changes in phase, since they are not perceived by the human eye, nor

by cameras. This problem can be solved by manipulating the light paths of the microscope such that the phase differences are converted into differences in amplitude (Frits Zernike, Section 8.1).

For phase contrast microscopy, a special condenser with an iris diaphragm and a phase shift ring is required. The phase shift ring, located in front of the objective before the plane of focus, is tasked with evening out the brightness of scattered and unaffected light, since the light that penetrated through the sample has decreased in intensity. In contrast to a conventional light field image, the background of a phase contrast picture is therefore dark. In addition, the phase shift ring amplifies the phase shift. Interference between scattered and unaffected light results in peaks and valleys that cancel out part of the light. The disadvantage of this increased phase shift is that when samples increase above a certain thickness (typically around the size of cell nuclei) halo effects can be seen.

Dark Field Microscopy A requirement for conducting dark field microscopy is the use of a special condenser, whose aperture is so large that the light beams that emerge from it pass by the objective. Only when a sample interrupts the light beam can diffracted light reach the objective. The imaged structures appear light on a dark background.

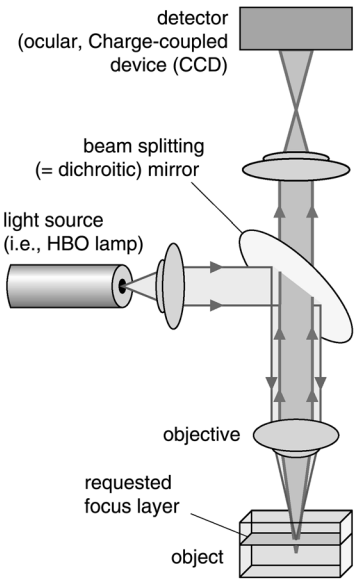


Figure 8.5 Principles of a fluorescence microscope.

Fluorescence Microscopy The focus here is on the function of a fluorescence microscope (Figure 8.5). A requirement for the use of fluorescence microscopic imaging is the use of fluorophores, which re-emit a portion of their absorbed light in the form of light of a longer wavelength.

A fluorescence microscope must have a strong light source (e.g., HBO lamp or laser, Section 8.4), an excitation filter, and an emission filter. The excitation filter is mounted below the sample in the light path. It is responsible for ensuring that only the characteristic, excitatory wavelengths for the chosen fluorophore reach the sample. The emission filter is placed between the objective and the ocular in the light path. There are fundamentally two different types:

- *Epifluorescence microscope* This type is by far the more commonly used for fluorescent microscopic imaging than the transmitted fluorescence microscope. In epifluorescence the objective functions as the condenser at the same time. The core of the epifluorescence is a special construction in the light path which brings in the excitatory beam of light. The excitation filter, the dichroic mirror, and the emission filter are located between the ocular and the objective.
- *Transmitted fluorescence microscope* This type is the older construction, which, in contrast to epifluorescence techniques, is seldom used and even then usually only with weakly magnifying objectives.

Polarization Microscopy Normally light vibrates in all directions. With the aid of special polarization filters, a particular plane of vibration can be filtered out such that linearly polarized light results. A second polarization filter, rotated to be perpendicular to the first, will filter this polarized light out entirely. Such rotatable polarization filters can be mounted in the light path of a microscope. The first polarization filter (polarizer) is below the condenser, the second filter (analyzer) is positioned above the objective. The use of polarizing microscopes is only worthwhile when the samples to be examined have polarizing characteristics. The main use of this procedure is therefore in mineralogy. In the area of bioanalytical research, this technique is rarely used and is mainly restricted to the field of botany (e.g., examination of the structure of starch granules or cellulose fibrils).

Interference Contrast Microscopy Differential interference contrast microscopy was developed based on polarization microscopy. Besides the polarizer and the analyzer of a polarization microscope, two Wollaston prisms, made up of two calcite prisms glued together, are needed. A polarized light beam is divided into two perpendicular beams at the interface between the two prisms. The first Wollaston prism is used in the front focus plane of the condenser, the second in the rear focus range of the objective. The object is illuminated by two beams whose polarization is perpendicular to one another. These are then phase shifted, depending on the thickness or index of refraction of the sample. Optimal interference contrast is observed on the edges of the sample, where the two partial beams are phase shifted differentially. In this case the orientation of the

sample is by no means unimportant. A rotating stage is usually employed to examine the sample from all sides. The second Wollaston prism reunites the two light beams. To generate interference, the planes of vibration must overlap, which is accomplished by the analyzer. An interference contrast image appears as a relief, which can lead to the illusion that it is a three-dimensional image of the sample structure. This is, however, not the case, since differences in density in the sample are translated into height differences in the image. In contrast to phase contrast microscopy, relatively thick samples can be visualized.

8.4 Detection Methods

The best method to stain or label a sample depends on the question at hand. Entire groups of cells in a tissue, specific cells within a group of cells, certain compartments within a cell, or individual molecules within a cell can be visualized selectively with the right stain or label. Histological staining mainly makes use of chemical or physicochemical stains.

To achieve a high specificity at the molecular level, a series of selective detection methods have been developed, particularly for proteins or other macromolecules. To increase the sensitivity of detection of macromolecules, the catalytic activity of intracellular enzymes is often employed. After the addition of suitable substrate molecules, each enzyme molecule produces many molecules of a visible reaction product. An even more sensitive detection method of macromolecules is the use of fluorophores.

Histological Staining The staining of cells and tissues is necessary for most purposes, since they consist of around 70% water. As a result, there are few structures present that guarantee sufficient contrast by refracting or absorbing the light beams in light microscopy. With the aid of various stains, intracellular and extracellular cell components can be colored selectively. For histological staining, paraffin slices are usually created, loaded onto slides and the paraffin removed with xylol and rehydrated with a series of alcohol baths of decreasing alcohol concentration to bring the sample back into an aqueous environment. The chemical basis for the specificity of many stains is unknown. Table 8.1 presents a few examples of histological stains.

Physical Staining The densest tissue is stained most intensively using immersion stains, while the method of stain absorption by solubility depends on the lipophilicity of the tissues. These stains dissolve more readily in lipid tissues than in the alcoholic solutions used and thus preferentially diffuse into fat tissue.

Physical Chemistry of Staining (Electro-adsorption) The physical basis of staining is the amphoteric nature of proteins. If the pH value is higher than the isoelectric point (IP), the structure has acidic groups and tends to form salts with basic stains and vice versa. In end point staining, in which the staining result is independent of its duration past a certain point, the differential IP of the cell nucleus and cell plasma is exploited. If a basic stain is present in acidic solution at a pH value between the IP of the nucleus (around 3.8) and that of the plasma (around 6.5) at around 4.5, the cell nuclei are selectively stained, since only the nuclear proteins have negative charges under these conditions.

Chemical Staining The reaction between stain and substrate takes place in accord with chemical laws. They allow detection of substances in a chemical sense. If the substance in question is not present, the reaction does not take place (e.g., iron detection).

Fluorescence Labeling Detection of macromolecules with the aid of fluorescence labeling uses the ability of fluorescing dyes (synonyms are fluorophore, fluorochrome) to absorb light of a particular wavelength and emit light at a longer wavelength. The excitation of a fluorophore with light in its absorption spectrum and subsequent microscopic examination through a filter, which only allows light of its emitted wavelength to pass, allows the light from the fluorophore to appear on top of a dark background.

Fluorophores are primarily used for the selective detection of macromolecules. In addition, they are used for the localization of enzyme activities in fluorometric variations of

Table 8.1 Examples of histologic stains.

Name	Staining results
Hematoxylin and eosin (HE)	blue: cell nucleus
(Figure 8.8)	red: cytoplasma, collagen fibers
Periodic acid–Schiff combined with	blue: cell nucleus
Hematoxylin (PAS)	purple: mucopolysaccharides, glycogen, fungi, parasites
Van Gieson (EvG)	black: elastin fibers
	red: collagen fibers
	yellow: muscle cells, fibrin, cytoplasma
Giemsa	blue: cell nucleus, basophilic cytoplasma
	red: collagen fibers, eosinophilic cytoplasma and granula
	violet: basal membranes, mast cell granula
May–Grünwald–Giemsa	blue: cell nucleus, basophilic substrates
	red: eosinophilic substrates, collagen fibers
Papanicolaou (Pap)	blue: cell nucleus
	blue–green: cytoplasma basophilic cells
	orange–red: ceratin
	pink: cytoplasma of acidophilic cells
Azan	blue: collagen fibers, reticular fibers
	red: cell nuclei, necrotic fibroids
Gomori silver	black: collagen fibers, reticular fibers
Nissl	pale blue: cytoplasma
	blue–violet: cell nucleus, tigroid
Congo red	red: amyloid
	blue: cell nucleus
Prussian blue	blue: iron
von Kossa	black: calcium phosphate
	red: cell nucleus

histochemical reactions, as well as the determination of turnover of cellular enzymatic reactions with fluorescent substrates.

With the combination of two or more fluorophores that differ in their spectral properties (absorption and emission characteristics), the co-localization of different subcellular structures or macromolecules can be examined. This is possible by the use of filters that are specifically tuned to the specific fluorophore. In addition, multiple fluorescence labeling is currently the only method that allows the direct investigation interactions between molecules and other components of the cell (Figure 8.6).

Direct and Indirect Immunofluorescence Labeling A common method to determine the localization of macromolecules in cells and tissues is immunofluorescent microscopy.

The detection of proteins is divided into direct and indirect immunofluorescent labeling. Direct immunofluorescent labeling involves the use of primary antibodies labeled with a fluorophore that bind directly to the macromolecule to be detected. Proteins can also be detected with fluorophore-labeled inhibitors and cofactors. The labeling of membranes can be performed with fluorescent compounds with lipophilic or amphiphilic characteristics, which are embedded in the membrane.

In indirect immunofluorescent labeling, a non-labeled primary antibody is detected by a specific, fluorophore-labeled secondary antibody. The advantage of indirect immunofluorescent labeling lies in the signal amplification, since multiple secondary antibodies can bind to a single

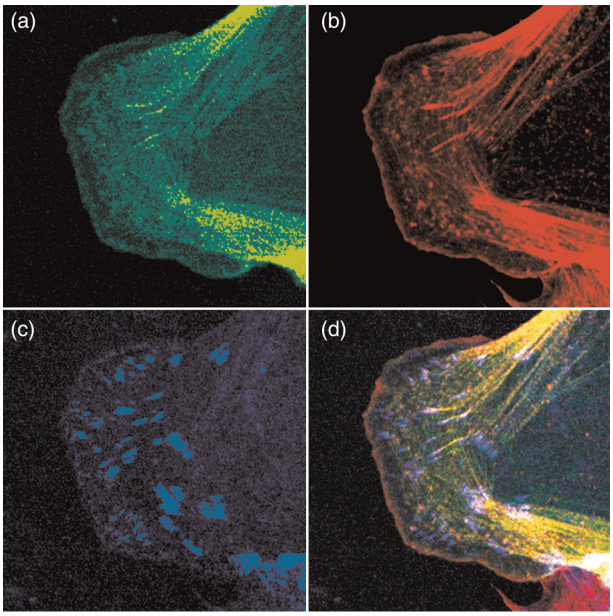


Figure 8.6 Triple fluorescence staining of a human keratinocyte. (a) GFP-Palladin (actin-binding protein), (b) AlexaFluor®-Phalloidin-labeled F-Actin, (c) antibody-staining of an cell adhesion molecule (Vinculin, Cy5-conjugated secondary antibody with fluorescence emission in infrared), (d) merged image of all three fluorescence signals. Source: courtesy of Dr. Bodo Borm, FZ Jülich, Germany.

primary antibody. Further advantages of indirect immunofluorescent labeling are the greater flexibility and the reduced costs in the combination of antibodies and fluorophores.

Immunofluorescent labeling is usually carried out using fixed material. To allow the antibodies to penetrate into intracellular compartments in fixed samples, permeabilization of the membranes with Triton X-100 or saponin is usually required. Unspecific binding sites are saturated with bovine serum albumin or specific sera depending on the question under investigation.

***In Vitro* Labeling with Organic Fluorophores** Despite the many advantages offered by *in vivo* labeling with fluorescent proteins (e.g., GFP fusion proteins), the known fluorescent proteins have the disadvantage that they are all relatively large, with a molecular weight of up to 27 kDa in their monomeric forms. Although there is great interest in finding smaller fluorescent proteins, all microbiological attempts so far have remained unsuccessful. Small organic fluorophores (e.g., fluorescein, rhodamine, carbocyanine) with a molecular weight of around 1 kDa represent an alternative to fluorescent proteins. With the aid of special, well-established affinity cytochemical (e.g., labeling of F-actin with fluorophore-labeled phalloidin) and immunocytochemical techniques, it is possible to achieve a specific subcellular labeling of molecules in fixed and permeabilized cells. The problem of possible steric hindrance and the resulting impact on protein function is significantly reduced when using small organic molecules. Besides which, changes in the local environment or distances or interactions between labeled sections of a protein can be detected by means of FRET (Förster resonance energy transfer).

GFP as a Unique Fluorescent Probe, Section 7.3.4

FRET, Sections 8.6 and 16.7

Fluorescence Labeling for Live Cell Imaging The increasing use of live cell imaging in bioanalytical research is thanks in particular to the newest developments using gene technological methods for the fluorescent labeling of proteins. Taking into account the physiological characteristics of cell is crucial for the establishment of new technologies in this area. Live cell imaging, and the qualitative and quantitative analysis that goes with it, require loading subcellular structures with fluorophore-labeled proteins. This requires striking the right balance between an efficient and stable fluorescent signal with a good signal to noise ratio, on the one hand, and the minimum possible interference with normal physiological processes, on the other. In particular the cellular localization and the physiological function of the protein under study should remain unchanged. Both the labeling and the subsequent imaging (influence of the light source, e.g. laser) can cause artefacts. Possible phototoxic reactions as a result of on the interaction of fluorophores or components of the culture medium or buffer with the laser light are also possible.

***In Vivo* Labeling with Organic Fluorophores** Two innovative technologies were recently developed that enable the labeling of specific recombinant proteins with small organic fluorophores in living cells (*in vivo*).

In the first method, recombinant proteins with a tetracysteine domain are labeled by adding fluorescein derivatives to the extracellular medium. These small, membrane-porous ligands only fluoresce when they bind with high specificity and affinity to cysteine residues.

In the second method, derivatives of human O6-alkylguanine-DNA-alkyltransferase (hAGT), a DNA repair protein, are used to bind covalently to fusion proteins. Intracellular esterases hydrolyze the acetate group of the fluorescein coupled to the hAGT derivative and activate its intracellular fluorescence. Other developments use the selective binding properties of chemical ligands to their corresponding receptor proteins. An example is the investigation of the regulation of the pH value of the various subcellular compartments during the course of secretion.

Labeling with Quantum Dots Besides small organic fluorophores, nanocrystals from the semiconductor industry, so-called quantum or Q-dots, are used as new fluorophores. The particles have a core diameter of 2–10 nm, roughly the size of a typical protein. They have a high photostability and absorb and emit over a wide range of wavelengths, which makes them well suited to the purpose. Despite these advantages relative to organic fluorophores and fluorescing proteins, their lack of biocompatibility has severely limited their use. New developments in the area of chemical coating (e.g., with streptavidin) have recently led to progress in solving this problem so that the imaging of multifluorescent labeling is possible over an extended period of time in living cells.

Quantum Dots, Section 7.3.5

***In Vivo* Labeling with Fluorescent Fusion Proteins (GFP and Variants)** Variants of green fluorescent protein (GFP) are frequently used, particularly for the examination of intracellular localization and the dynamics of proteins in living cells. In the middle of the last century, GFP was discovered accidentally during the purification of aequorin from the luminescent jellyfish *Aequorea victoria* (Figure 8.7a). At this time the ability of the protein to fluoresce an intense green under UV light was already apparent. By fusion and subsequent expression of the GFP gene with the gene of interest, the intracellular localization and dynamics of the resulting protein can be examined using fluorescent microscopic analysis. Since GFP fluoresces spontaneously when fused to another protein, it can be expressed in cells *in vivo* by gene transfer. The alternative, to heterologously express, purify, label, and micro-inject recombinant proteins, is far more laborious, expensive, and time consuming. The Nobel Prize in Chemistry in 2008 was awarded for the discovery of GFP and the development of genetic labels (Section 8.1).

GFP, Section 7.3.4

By mutating GFP from *Aequorea* it was possible to generate spectral variants with blue and yellow–green emissions (BFP, CFP, YFP); however, no variants whose emission wavelength maxima is over 529 nm have yet been found. Fortunately, the discovery of new types of GFP-like proteins in Anthozoa (corals) was able to expand the spectrum by almost 30 significantly different fluorophores (Table 8.2). Despite only a low sequence similarity of the different GFP-like proteins, all of them have an eleven beta barrel structure (Figure 8.7b), which is threaded along its axis by an alpha helix. The chromophore is part of this alpha helix and is located in the middle of the beta barrel. The chromophore is a *p*-hydroxybenzylidene-imidazolidine, which consists of the cyclic tripeptide Ser65, Tyr66, and Gly67 in the wild-type GFP (wtGFP, Figure 8.7c).

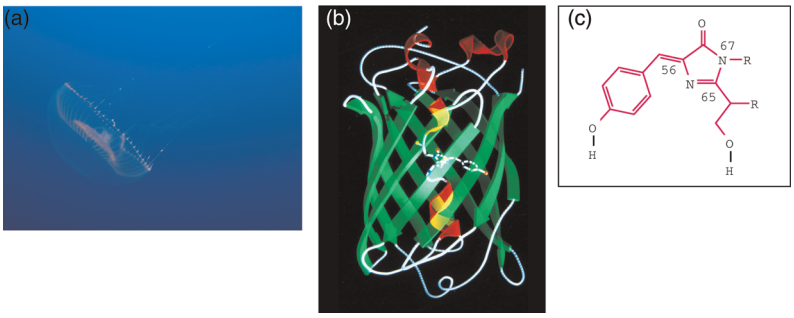


Figure 8.7 *Aequorea victoria,* a bioluminescent hydrozoan jellyfish Source: Monterey Bay Aquarium, CA, USA; Mnolf, CC-SA-3.0 (a). Structure of green fluorescent protein (GFP) and localization of chromophore Source: K. Brejc, T.K. Sixma, P.A. Kittis et al. *Proc. Natl. Acad. Sci. USA* 1997, 94, pp. 2306–2311, with permission © 1997 PNAS. (b). Chromophore of wildtype GFP, composed of cyclic tripeptide Ser^{65}, Tyr^{66}, and Gly^{67} (c).

Table 8.2 GFP variants.

Species	Fluorescent protein	Mutants (alloforms)	Absorption maximum (nm)	Emission maximum (nm)
		wtGFP (green fluorescent protein)	395	475
Bioluminescent jelly fish *Aequorea victoria*	*Aequorea*-GFP	—		
		BFB (blue fluorescent protein)		
		CFP (cyan fluorescent protein)	383	445
		YFP (yellow fluorescent protein)	434	477
		PA-GFP (photoactivatable GFP)	514	527
Bioluminescent soft coral *Renilla reniformis*	*Renilla*-GFP	—	395	475
Non-bioluminescent coral species	DsRed (drFP583)	T1	558	583
		E57		
		E5		
		mRFP1		
	cgigCP	HcRed		
	EqFP611	—		
	AsFP595	KFP1		
	Kaede	—		

GFP is, strictly speaking, the only fluorescent protein ever discovered, since the chromophore actually forms a portion of the peptide chain. Its spontaneous fluorescence is based on the autocatalytic synthesis of the chromophore. Although most GFP-like proteins belong to the group of fluorescing proteins, there are a few of them that show a strong absorption but no emission; these are referred to as chromoproteins. Other known proteins serve as fluorophores, but without exception the actual chromophore is bound as a cofactor, such as, for example, lumazine or flavine.

Fluorophores and Light Sources for Fluorescence Microscopy The range of light visible to the human eye is from 400 nm (violet) to 700 nm (red). A few stains that are used in biological detection applications absorb light in the UV range (Table 8.3). As a result, a suitable light source must also cover this range. A common UV fluorescent dye is Fura-2, which is used for the visualization of calcium signals.

The fluorescence intensity of labeled biological preparations is, in comparison to the intensity of the excitation light, usually quite low. Loading of the sample with fluorophores is not usually the limiting factor, instead it is primarily determined by the efficiency with which the dye absorbs and emits photons. Furthermore, the ability of the dye to go through repeated absorption/emission cycles plays an important role. The quantification of absorption and emission are calculated to give the coefficient of absorption ε for the absorption and the quantum yield (QY) for the fluorescence (this is the ratio between emitted and absorbed photons). Both constants are dependent on the conditions. The fluorescence intensity of a dye is proportional to the product of ε and QY.

Besides this, the intensity of the fluorescence is reduced by the optical properties of the microscope. Accordingly, only a portion of the emitted light, a certain cone of light, whose size depends on the numeric aperture of the objective used, is captured by the optics of the microscope and registered by the detector (eye, camera, photomultiplier). For these reasons it is understandable that for the creation of strong emission signals a strong light source is required. In addition, in

Table 8.3 Fluorophores and their spectral properties (selected).

Fluorophore		Absorption maximum (nm)	Emission maximum (nm)	Color
AMCA	Aminomethylcoumarin-acetate	350	450	Blue
DAPI		359	461	
Cy2	Carbocyanine	492	510	Yellow–green
DTAF	Dichlorotriazinylamino-fluorescein	492	520	
FITC	Fluorescein-isothiocyanate	492	520	
	AlexaFluor 488™	495, 492	519, 520	
	Fluo-3	506	526	
CFDA		494	520	
	Acridine orange + DNA	502	526	
Cy3	Indocarbocyanine	550	570	Orange–red
TRITC	Tetramethyl-rhodamine	550	570	
	AlexaFluor 546™	556, 557	572, 573	
R-PE	Phycoerythrin from *Porphyridium cruentum*	488, 565	585	
Rhod. Red-X	Rhodamine Red-X	570	590	
Texas Red	Rhodamine derivatives	596	620	
Cy5	Indodicarbocyanine	650	670	Deep-/infra-red
	Acridine orange + RNA	460	650	
APC	Allophycocyanin	650	660	

particular for time courses during live cell imaging, the light source must be stable over an extended period of time. This means that oscillation and flickering of the light intensity during the data acquisition process must be negligibly small. They should be considerably below the value of the expected changes in signal intensity during the course of the experiment.

Types of Light Sources Unfortunately, there are no light sources that cover the entire usable spectrum (near-UV to near-visible infrared). As a result, the light source must be selected according to the planned application. Besides lasers and diodes, there is also a difference between incandescent and non-incandescent lights.

Incandescent Lights Incandescent lights create light through the heating of filaments, usually tungsten, by means of electric energy. The filaments are either in a vacuum (standard incandescent bulbs) or in a noble gas (nitrogen or noble gas) or a mixture of noble and halogen gasses (halogen lights). Tungsten lights are not normally used in fluorescence microscopy because they are too weak in the UV and blue wavelength part of the spectrum. Non-incandescent bulbs without a filament are based on electric discharge inside a gas. When the electric potential between the two electrodes is large enough, the gas is ionized and electricity flows from one pole to the other. After the charge the electrons and gas ions rearrange themselves and release energy in the form of light. The gas in neon tubes glows with an orange–red color, while commonly available fluorescence mercury gas lamps produce ultraviolet light. The inside of the glass tube is coated with phosphorous, which absorbs this light and emits it again as bright white light. Most non-incandescent lamps used in microscopy are mercury or xenon gas lamps. They consist of two electrodes, which are sealed under high pressure inside quartz glass. When an electric charge is applied, electrons bridge the gap between the electrodes, exciting the electrons of the gas atoms, putting them into a higher energy state. This energy is released in the form of light when the atoms return to their ground state. Gas lamps must first be heated to reach their maximum intensity. After a certain amount of time, the lamps decline in efficiency and there is a chance of implosion.

Table 8.4 Emission wavelengths of lasers in comparison to HBO lamps.

Source	Emission wavelengths (nm)
Argon	351 364 458 466 477 488 496 502 514
Krypton	337 365 468 476 482 521 531 568 647
Argon/krypton	488 568 647
Helium/neon	543 594 604 612 629 633 1152
Helium/cadmium	325 442 534 539 636
HBO lamps	313 334 365 405 436 546 577

Mercury Vapor Lamps (HBO Lamps) Mercury vapor lamps are common light sources used in microscopy. One disadvantage of these lamps is that they produce a very uneven emission spectrum, with distinct peaks in the near-UV (365 nm), violet (406 nm), blue (435 nm), green (546 nm), and yellow (578 nm) wavelengths. At other wavelengths their emissions are more even, but not as intense. In particular, it is unfortunate that they lack a peak in the range of 480 nm, where common fluorophores, such as Cy2, AlexaFluor 488, FITC, and GFP, absorb best.

Xenon Vapor Lamps Xenon vapor lamps are frequently used for fluorescence microscopy because they produce a smoother emission spectrum throughout the entire visible range than mercury vapor lamps. However, the intensity of this type of lamp also decreases in the near-UV range, but remains sufficient for most UV dyes. Due to this advantage, xenon lamps are favored over mercury vapor lamps, for example, when using CCD cameras. The strong peak in the infrared range indicates that a significant portion of the energy is emitted in the form of heat. Regulation of the accumulation of heat in a xenon lamp housing is of critical importance. Temperature variation, which can be caused, for example, by uneven airflow, has a direct impact on the light intensity and therefore interferes with quantitative analysis.

Mercury–Xenon Vapor Lamps These relatively new lamps contain a mixture of the two gases and thereby combine the best properties of each. The lines in the UV range are, in comparison to conventional mercury vapor lamps, stronger and sharper. Of greater importance for microscopic applications is the lower variation in light intensity and the significantly longer lifetime of the lamps.

Lasers Lasers are, due to their uniform emission spectrum, better for the specific generation of the proper excitation wavelength for a particular fluorophore. Every laser emits in a characteristic wavelength range, so that the detectable fluorophores are therefore predetermined. Table 8.4 shows the available emission wavelengths of the most common lasers, and in comparison the strongest peak in the spectrum of an HBO lamp. It is characteristic that the peaks of the HBO lamps range further than the spectral lines of the laser and that, in addition, a significant emission exists between the peaks. As a result, the choice of usable fluorophores that can be excited with a HBO lamp is greater than that of any one laser.

8.5 Sample Preparation

With respect to sample preparation for light microscopic investigations, there are two fundamentally different areas of application: first, the visualization of fixed materials and, second, live cell imaging of living samples. Furthermore, the sample preparation methods differ between single cells and tissue biopsies.

Isolated Cells For the microscopic investigation of isolated cells, the cells are usually seeded onto thin glass materials (0.15 mm thick) in multiwell plates for tissue culture. For fluorescence microscopy of fixed, adherent cells, cover slips precoated with poly-L-lysine are usually used. The positive charge of the amino acid allows better adherence to the surface of the negatively charged cells (glycoproteins and carbohydrates). For microscopic investigation of the interaction of cells

with components of the extracellular matrix, for example, during adhesion or migration, the cover slips are coated with the corresponding components (e.g., fibronectin, collagen, laminin, and others). For fluorescence microscopy of fixed suspension cells, the preparation of a smear of suspension is often recommended to allow better attachment of the cells.

For live cell imaging, cells are seeded in glass bottomed wells, which, depending on the question addressed, are either self-made or chosen from the wide selection of commercial offerings.

Tissue Biopsies Aside from intravital and multiphoton microscopy, the microscopic investigation of tissue biopsies requires the aid of special microtomes to prepare shock frozen paraffin-embedded slices. The slices are mounted on slides and subsequently treated with dyes or detection solutions in order to detect the desired structures.

Fixation Tissue autolysis begins immediately after biopsy. The tissue sample must, therefore, be fixed or prepared immediately after removal of the biopsy. Fixing stops the autolysis. The most commonly used fixing agent is formalin, a diluted and buffered formaldehyde solution, which gives the tissue a rubber-like consistency. Formaldehyde crosslinks proteins, by binding to free amino groups.

The disadvantage of formaldehyde fixation is that the proteins lose, under certain conditions, their functions and antigenic characteristics. Therefore, many enzyme histochemical and immunocytochemical experiments cannot be carried out on formaldehyde fixed material. In the aqueous formalin solution, glycogen and many other polysaccharides, as well as certain crystals such as urate, are dissolved and washed away.

For electron microscopic investigations, glutaraldehyde is often used, which, as a result of its two aldehyde groups, is more reactive than formaldehyde. Fixing with glutaraldehyde is not suitable for fluorescence microscopic applications, since the double bond it contains can lead to autofluorescence.

Rather than aldehyde, fixation can be carried out with organic solvents, such as methanol or acetone, which involves their ability to dehydrate the samples. Frequently, organic solvent-based fixation preserves the antigenicity better, at the cost, however, of preservation of the ultrastructure of the sample. The most suitable fixing material needs to be experimentally determined for each application. The duration of fixing depends on the size of the sample. The fixation agent must be present at a significant excess with respect to the volume of the sample. For large samples it may be necessary to dissect the biopsy material to guarantee an even penetration of the fixation agent.

For pathological diagnostics, which may arise quickly as slices (and/or impression smears) during the course of an ongoing operation, it is necessary to work with fresh, unfixed biopsy samples to save time.

In addition, many enzyme histochemical tests can only be carried out on unfixed tissues. In immunohistology there is a much broader range of antibodies available for unfixed cryostat slices than for fixed tissues. Furthermore, DNA and RNA can be isolated from fresh tissue for molecular pathological diagnostics. When needed, tissue can be used for microbiological diagnostics.

Paraffin Samples These samples are naturally suited for the histological analysis of tissue biopsies (Figure 8.8). Immunofluorescence labeling works better, however, on frozen slices.

Embedding The preparation of histological thin slices requires that the biopsy material is stable and has an even consistency. This is accomplished by embedding the tissue in paraffin. The biopsy sample must be oriented in the embedding cassette such that the sliced samples have the desired orientation (e.g., cross section of the skin). In pathological diagnostics the correct orientation of the biopsy material is necessary to determine the extent of a tumor, the areas to be removed, and the degree of metastasis. For bones or tooth samples, before embedding in paraffin the tissue must be decalcified. For untreated hard tissues (teeth, bones), instead of paraffin, acrylic or plastic is used, which allows the preparation of slices or ground down samples without decalcification.

The biopsy sample is infiltrated with hot paraffin wax, which solidifies as it cools. Since paraffin is not water soluble, the tissue must first be dehydrated in a series of alcohol baths of increasing alcohol concentration. Subsequently, the alcohol is replaced by an interim medium, typically xylene, and finally replaced by paraffin wax. The paraffin saturated tissue samples are

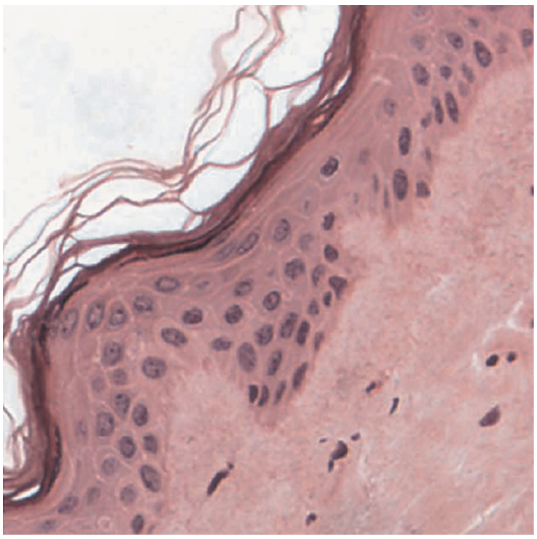


Figure 8.8 Section of paraffin-embedded human skin following haematoxylin and eosin staining. Blue: nuclei, red: cytoplasm, collagen fiber. Source: Dr. T. Quast, University of Bonn, Germany.

then laid in a mold, filled with paraffin and allowed to solidify into a paraffin block. After solidifying, the block is removed from the mold.

Creation of Paraffin Slices Tissue slices several micrometers thick (4–8 µm) can only be made using special devices. For the preparation of paraffin slices, rotary or sliding microtomes with special steel or disposable blades are used.

Frozen Slices

Embedding The biopsy sample is infiltrated in a special sugar and antifreeze-containing embedding medium, which solidifies at temperatures below −20 °C. Embedding usually takes place in embedding cassettes over an atmosphere of liquid nitrogen. Storage of embedded frozen samples should take place at −80 °C. Unfixed tissue can be solidified by freezing and, therefore, be made sliceable. This method is used when speed of diagnosis is of the essence, such as during the course of an ongoing operation. However, the freezing and increased mechanical demands of the unfixed tissue leads to artefacts more frequently. In addition, there is a greater danger of infection associated with unfixed tissue samples. The frozen slice method is particularly suitable for the creation of slices in which fat tissues are represented, since in this case there is no need for fat dissolving substances, and for the creation of enzyme histochemical and immunohistological staining.

Creation of Frozen Slices (Rapid Slices) For the preparation of frozen slices, refrigerated cryomicrotomes (cryostats) are used. Otherwise the procedure is comparable to the creation of paraffin slices (see above).

Sealing For microscopic examination and storage, the stained preparations are conserved by sealing them with a special sealing medium between the cover slip and the slide. This procedure can also be done automatically with the aid of sealing machines. The sealing medium should keep the preparation transparent and at the same time not damage the structures and staining. For fluorescent preparations, a non-fluorescent sealing medium should be used that, in addition, contains protection against bleaching, to reduce the bleaching during microscopy.

8.6 Special Fluorescence Microscopic Analysis

Highly developed fluorescence microscopy has made it possible to investigate the localization, dynamics, and interactions of molecules in living cells (live cell imaging). In conventional fluorescence microscopy, the resolution of the z-plane is limited by the fact that emission signals above and below the plane of focus are also detected (Figure 8.9).

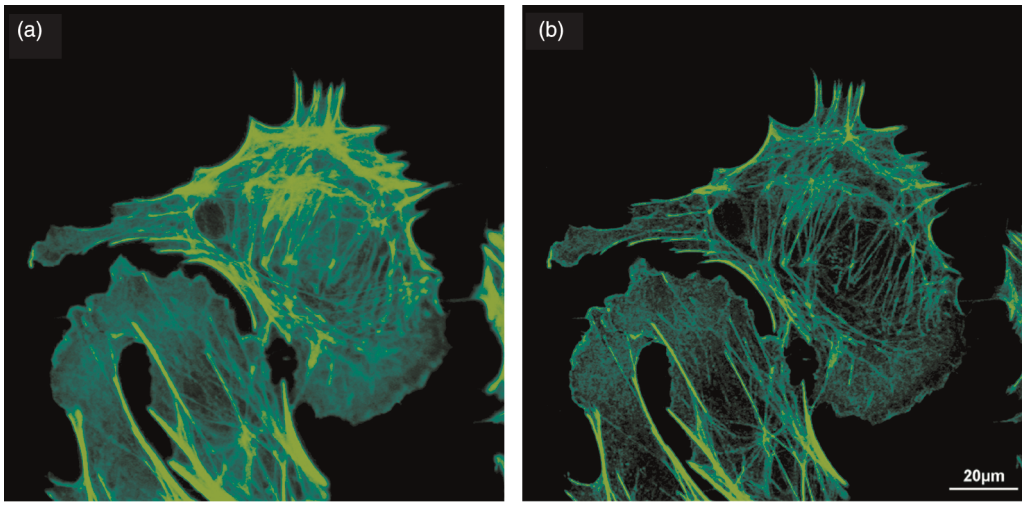


Figure 8.9 Comparison of confocal and epi-fluorescence light microscopy. FITC-phalloidin staining of F-actin in human fibroblasts. (a) Non-confocal epifluorescence, (b) confocal laser scanning microscopy. Source: By courtesy of Dr. Bodo Borm, FZ Jülich, Germany.

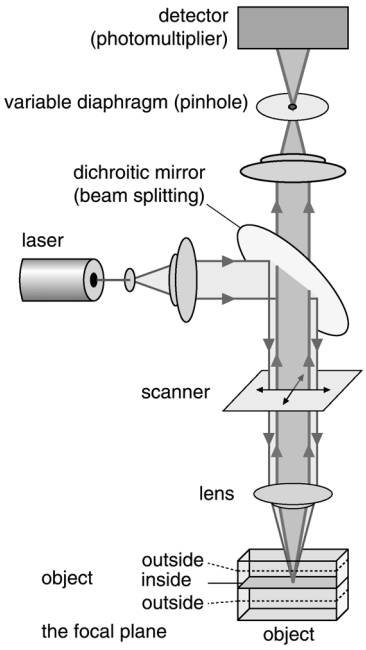


Figure 8.10 Principle of confocal laser scanning microscopy (cLSM).

With the aid of confocal LSM (Figure 8.10) it is possible to gather only the emitted light from the plane of focus of a sample. This involves focusing the bright light of a laser over the slide on a point in the sample in the desired plane. The light emitted from there is focused on a variable diaphragm pinhole and recorded by a detector behind the pinhole, typically a photomultiplier. Since the point of focus and the pinhole lie in conjugate focus planes, and therefore are confocal, only light from the focus point can pass through the pinhole. Diffracted light emitted above or below the plane of focus is effectively blocked by the confocal pinhole (Figure 8.9b). Analogous to a conventional fluorescence microscope (Section 8.3), LSM also has a dichroic (beam-splitting) mirror, which reflects excitatory light of a particular wavelength onto the sample and directs the light emitted by the sample to the detector. New LSM systems employ an AOBS (acousto-optical beam-splitter) instead of a dichroitic mirror, which has the advantage of a greater passage of laser light to the sample and greater flexibility of detection.

In contrast to conventional fluorescence microscopy, which involves excitation and detection of the entire field of view, excitation and detection using LSM is performed as a scan, point for point, line for line.

With the aid of galvanometric scanning mirrors, the laser light is directed sequentially through the sample to give a contrast-rich, high resolution optical slice. Through the step-wise shift of the focal plane, a stack of optical slices in the z plane can be prepared, which can subsequently be digitally processed into three-dimensional reconstructions. Although the procedures to create the images are significantly different, confocal optical slices of cells can be seen as comparable to computer tomographic pictures of tissues in the medical field.

LSM allows the spatial and temporal high resolution of morphologic details of groups of cells in tissues, as well as on the subcellular level. With these imaging procedures, physiological and pathological correlations can be analyzed. As a result of the slow rate of image generation (low number of pictures per second), conventional LSM systems are not suitable for live cell imaging of fast cellular processes. On the other hand, relative to Nipkow systems, these systems have the advantage that excitation can be limited to a small section of the sample, without which it is impossible to carry out photobleaching experiments (e.g., FRAP, see below and Table 8.5).

Multiphoton Fluorescence Microscopy Multiphoton microscopy is a modification of confocal LSM that enables the creation of high resolution images of biological preparations using nonlinear optical processes.

A basic characteristic of fluorescence is that the excitation light must be at least as energetic as the light emitted by the molecule. It is, therefore, not possible to excite a blue fluorescing molecule (high energy) with a photon of red light (low energy). From the perspective of a microscope user, however, it is desirable to use light of the lowest possible energy, since it reduces the chance of damage to the tissue caused by the light. The solution to this problem is

two-photon absorption using high intensity infrared light. In this case, a molecule absorbs two photons almost simultaneously and thereby absorbs twice the energy of a single photon.

The high light intensity required for two-photon absorption is generated with a pulsed laser. They emit the light energy in extremely short pulses, which results in peak pulses of up to 100 kW. The repeat frequency of the pulses is between 10 and 100 MHz. Titanium sapphire femtosecond lasers are the most common type. They are, however, difficult to adjust and react sensitively to long term drifts in their complex mechanics. The high costs associated with the procurement, running, and maintenance prevent the widespread adoption of two-photon microscopy as a routine procedure. The light intensity decreases outside the focus area as a quadratic function in two-photon absorption, just as with conventional confocal LSM, so that a high local resolution in the z plane is guaranteed. In addition, the probability of two-photon absorption decreases with the square of the light intensity, so that for two-photon microscopy the dependence on the light intensity on the point of focus is a power of four. The result is a greatly improved localization in the z plane, as a result of the reduction of the so-called background fluorescence from areas outside the focus plane.

The improved localization results in the following advantages: photobleaching and the damaging influence of the excitation light on the sample are limited to a very small area, while other areas remain untouched and available for further investigation. Owing to the quadratic dependence of the two-photon absorption, the pinhole in front of the detector can optionally be omitted.

The wavelengths of the excitation light and emission signals are far apart, so that the fluorescence can be observed over a broad spectral range, without leading to overlap with the excitation light. This results is a very low background signal with high fluorescent signals. Pulsed laser light from the near-infrared part of the spectrum is only weakly diffracted and absorbed by most biological substances, so that it can penetrate far into the tissue. Therefore two-photon microscopy is well suited for the investigation of tissue biopsies. Infrared light is less damaging to cells than visible light or UV radiation and allows examination even of light sensitive structures, such as retina cells.

Intravital microscopy is the observation of the interactions of endogenous or exogenous cells in an operatively accessible tissue (e.g., extravasation of leukocytes) while using multiphoton microscopy.

Confocal High-Speed-Spinning Disk Systems (Nipkow Systems) With extremely high imaging rates of up to several hundred images a second, these systems are primarily useful for the visualization of fast processes in real time (e.g., Ca^{2+} imaging and FRET, see below). Nipkow systems have two rapidly rotating disks: one is behind the laser light source, where it focuses the illumination with the aid of a large number of small lenses, and the other disk uses several thousand synchronously rotating pinholes to generate the confocality (Figure 8.11). In contrast to conventional LSM systems, the detection does not involve a photomultiplier, instead CCD (charge-coupled device) cameras are employed, which gather the emitted light of all the pinholes quickly and simultaneously. In this manner phototoxic effects and photobleaching is reduced.

Live Cell Imaging Microscopy in the biological sciences is no longer confined to the pure structural characterization of fixed tissue and cells, instead it also involves the investigation of dynamic processes in living cells with the aid of new fluorescent methods. An important clue to the function of a protein is given by its localization and dynamics inside the cell. Work on living cells is particularly revealing, since here the effects of substances (inhibitors, pharmaceuticals, and others) can be analyzed directly.

Dynamic processes like cell migration, cell growth, metabolic transport, and signal transduction can last less than a second or last several days. Accordingly the requirements for imaging vary from several pictures per second to intervals of many minutes. For short intervals, picture acquisition rate, light sensitivity, and exposure time are the limiting factors; for time lapse series, which span long periods of time, maintenance of the proper plane of focus and the maintenance of stable culture conditions (temperature, CO_2, humidity) are paramount.

Many components of microscopic systems limit the speed of image acquisition (e.g., the change of laser paths or filters). Systems that use monochrome cameras have the advantage that rapid switching between the excitation wavelengths is possible (often below 3 ms), but, on the other hand, they suffer from lowered illumination intensity, which is a result of coupling optical fibers into the microscope. Filter wheels usually allow a higher transparency to light, but take

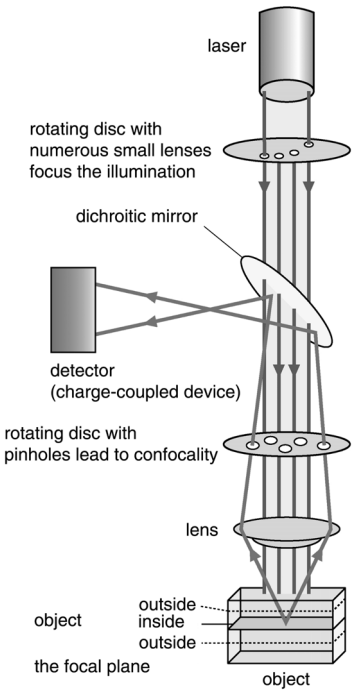


Figure 8.11 Principle of confocal spinning disk microscopy (Nipkow).

Table 8.5 Comparison of various systems for live cell imaging (summary).

	Conventional fluorescence microscopy (wide field)	Confocal *laser-scanning*-microscope	Confocal *spinning-disk* system (Nipkow)
Confocal	Not confocal	A variable *pinhole* eliminates emission signals outside the plane of focus	A rotating disk with small lenses focusses the excitation light
Principle	Emission signals above and below the plane of focus are detected		A second disk with synchronously rotating *pinholes* creates the confocality
Light source	Mercury vapor lamps or mercury xenon vapor-lamps	Laser	Laser
Detector	CCD camera	Photomultiplier	CCD camera
Advantages	– High rate of image acquisition, flexibility with respect to wavelength of excitation and emission, filter sets available	– The excitation can be specifically focused on a small section of the sample (requirement for photobleaching experiments)	– High rate of image acquisition with the use of a laser beam
			– Reduced phototoxicity due to decreased bleaching
	– Rapid change of filters with the use of filter wheels or monochromators is possible		
Disadvantages	– Not confocal	– Photomultipliers are usually less sensitive than CCD	– Limited rate of image acquisition due to delays
		– The scanning rate limits image acquisition rates	– Photobleaching experiments not possible
		– Switching laser paths	

significantly longer to switch. Scanning systems are slower than CCD cameras, since they do not capture the whole image at once, but instead acquire the data pixel by pixel.

The choice of a suitable system for the visualization of living cells is mostly dependent on three factors:

- the sensitivity of the detector,
- the image acquisition rate,
- the sample preservation, avoiding the disruption of normal physiological processes.

The light microscopic observation of living preparations requires consideration of both a good signal to background ratio and also potential damage to the preparation caused by overexposure to light. Advanced live cell imaging required the development of methods that minimize phototoxicity due to light interactions (e.g., with laser light) during imaging, as well as allowing the survival of the cells under stable culture conditions. Fully automatic and interactive microscope control is possible with the aid of sophisticated software. Software also allows complex protocols to be configured easily. Configuration parameters can be saved and reused in later experiments. The imaging methods based on fluorescence detection have been described above. The overview given in Table 8.5 summarizes the important characteristics of these techniques and their significance for live cell imaging.

Light Microscopic Super Resolution below the Abbe Limit Light microscopic techniques are important methods of modern cell biology and are, in combination with immunofluorescent markers, fluorescing fusion proteins, or *in situ* techniques, important tools for the specific localization of almost all cellular components. Aside from the significance for biological questions, all conventional methods of fluorescence microscopy, including wide field, laser scanning, and multiphoton microscopy, have a lower limit of resolution that is limited by the light refraction when passing through lenses and apertures. It has long been an important goal to improve the resolution of light microscopy below the optical diffraction limit of 200 nm, while retaining its advantages relative to electron microscopy, in particular in examining living cells. The wave-like nature of diffracted light prevents the visualization of preparations that are smaller than 200 nm in the lateral (x, y) directions and smaller than 500 nm in the axial (z) direction. Since

most subcellular structures, such as cytoskeletal filaments and vesicles, are significantly smaller, various techniques have been in development since the 1990s that allow a resolution below the Abbe limit (Section 8.3), resulting in "super" resolution. A well-developed technique is based on reducing the size of the focused light spot using nonlinear methods. Examples of microscopic techniques using super resolution are STED (stimulated emission depletion), GSD (ground state depletion), and SSIM (saturated structured illumination). These techniques allow a resolution of between 20 and 50 nm in the lateral plane by employing a specific reduction of the point spread function (PSF). Stochastic methods for super-resolution imaging rely on a temporal confinement of the fluorescence signal and the detection of single fluorophores. Examples of such techniques are PALM (photoactivated localization microscopy) and STORM (stochastic optical reconstruction microscopy). Other techniques do not involve the focusing of light at all, instead relying on the characteristics of evanescent waves, such as TIRFM (total internal reflection fluorescence microscopy), or the use of small nano-apertures as light guides as in SNOM/NSOM (near-field scanning optical microscopy).

Electron Microscopy, Chapter 19

STED (Stimulated Emission Depletion) STED microscopy goes below the limit posed by light diffraction by the targeted deactivation of fluorescent dyes. The area of fluorescence emitted from the sample is then much smaller than the area excited by the laser beam. This is made possible by targeted depletion of the fluorescent molecules in the edge regions of the focus. This is accomplished by employing a second laser – in addition to the one that excites the focus area in conventional confocal LSM – that deactivates the fluorescent molecules in the fringe area around the central focus point. This second laser has a ring-shaped profile in the focus plane. The second laser is completely dark where the central laser's excitation is maximal. The second, depletion laser beam has no influence on the fluorescent molecules in the focus area of the primary excitation laser beam. It bleaches out the fluorescent molecules in the fringe areas of the area of focus by stimulated emission and as a result the molecules in the fringe area remain dark, even though they are in the area excited by the primary laser. Therefore, only the fluorescent molecules exactly in center of the focus area are excited. If the depletion laser is of high intensity, this area is much smaller than the area hit by the excitation laser. The detection of the sample measures light emanating from a much smaller spot than a conventional confocal LSM, which allows the resolution to be higher.

PALM (Photoactivated Localization Microscopy) and STORM (Stochastic Optical Reconstruction Microscopy) PALM and STORM are special procedures to improve the resolution of a light microscope in which one uses the special characteristics of photoactivatable (PA) proteins or photoswitchable synthetic fluorophores. PA proteins are variants of GFP that can be specifically activated and deactivated with light of specific wavelengths. A short pulse of light randomly activates a few of these inactive, non-fluorescing PA proteins to begin fluorescing. Continuing excitation leads to irreversible bleaching of the PA molecules. During the bleaching process, continuous imaging and position determination of the molecules is carried out. The conditions are chosen such that the probability of activation of two nearby PA molecules is small. This procedure is repeated until all the PA proteins are bleached. Along with PA proteins, organic fluorophores can be designed to become photoswitchable and used in a similar fashion. At first the fluorescing molecules cannot be resolved due to the usual limits posed by light diffraction. The concept of photoactivation or photoswitching, however, allows us to separate the detection of fluorophores in time. The rate of photoactivation is tuned such that, at a given time, only a subset of fluorophores is activated and detected as single emitters. Through the use of special algorithms that employs a mathematical function to approximate the point spread function (see microscopy glossary in Section 8.3), the exact position of the spatially isolated molecules is determined and processed into the final high resolution picture.

(S)SIM ((Saturated) Structured Illumination Microscopy) SIM employs excitation with intensity-patterned light. The light pattern is shifted and rotated, and several (typically five for 2D SIM) images of the specimen are recorded. Image processing retrieves high-frequency information from the raw data, and provides a reconstructed image with a lateral resolution better by up to a factor of two compared to the resolution limit. 3D SIM enables superior spatial resolution below 200 nm along the axial coordinate through the use of structured illumination provided by three overlapping light beams. SSIM is an extension of SIM by introducing nonlinearity arising from saturation of fluorescence, and can achieve a spatial resolution of 50 nm or better.

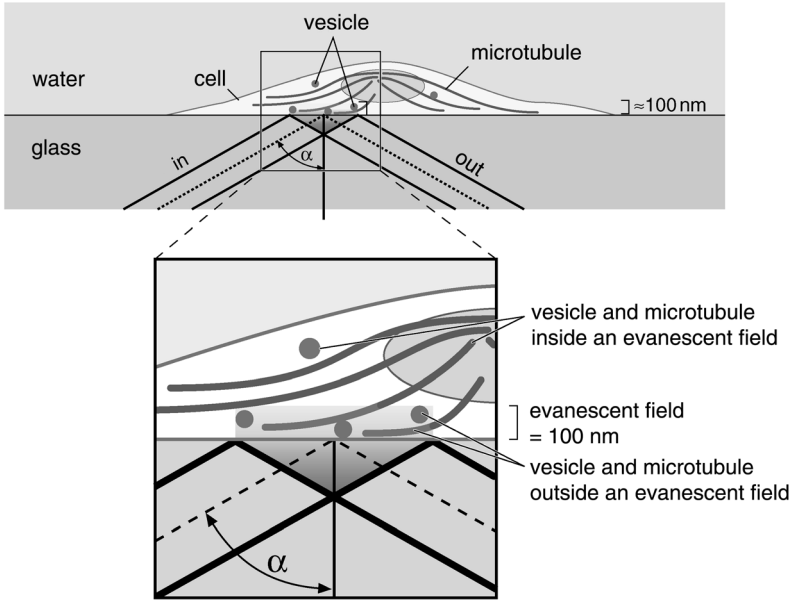


Figure 8.12 Total internal reflection microscopy (TIRFM). Total reflection happens when the angle of incidence of the excitation light is greater than a critical angle. A so-called evanescent field results when a medium with a high index of diffraction (here: cover glass) passes into a medium with a lower index of diffraction (here: cells). The intensity of the evanescent field decreases as the distance from the cover slip increases so that only an area of about 100 nm of the basal cells bordering the cover slip are excited. As a result, only the fluorescent signal coming from the area near the plasma membrane is visualized and not from areas in other planes of focus. TIRFM can be combined with conventional wide field microscopy to visualize events outside of the evanescent field. Adapted from: Stephens and Allan (2003) *Science*, **300**, 82–86 (see Further Reading).

TIRFM (Total Internal Reflection Fluorescence Microscopy) The variant TIRFM is primarily useful for the visualization of cellular processes taking place in distinct regions of the cell or plasma membrane (Figure 8.12). It allows the detection of events in the immediate proximity of the cover slip. Laser excitation in a glass medium at a certain critical angle leads to total reflection. TIRFM takes advantage of the phenomena of evanescent fields, which result when a medium with a low refractive index is next to the glass. A portion of the incident light is not reflected but instead diffuses into the optically less dense cell. The thickness of the evanescent field is about 100 nm; the intensity drops exponentially as the distance from the cover slip increases. The result is that the emissions only result from this defined layer and background signals from the rest of the preparation are reduced. The coupling of laser light in special objectives makes rapid switching between TIRF and wide field illumination possible and, thus, rapid switching between observation of cell surface and intracellular structures.

NSOM/SNOM (Near-Field Scanning Optical Microscopy) NSOM/SNOM also uses the characteristics of evanescent fields, similar to TIRFM, to achieve high spectral resolution in both time and space. In contrast to TIRFM, the evanescent field is achieved by placing the detector in the immediate proximity of the preparation, such that the distance is significantly smaller than the wavelength of the light. In this case resolution is not limited by the wavelength of the light used, but instead by the aperture of the detector.

Measurement of the Movement of Molecules

FRAP (Fluorescence Recovery after Photobleaching) FRAP is an optical method for the measurement of diffusion or molecular movement in cells or thin liquid films (Figure 8.13). Initially, the intensity of fluorescently marked molecules in a specific area is measured. Next, a laser pulse is used to irreversibly bleach out the fluorescence of this area. The diffusion time can then be calculated by measuring the rate at which the fluorescing molecules of neighboring areas flow into the area under investigation.

FLIP (Fluorescence Loss in Photobleaching) FLIP is used similarly to FRAP to investigate the molecular dynamics in 4D (x, y, z, and t directions) in living cells, for example, the dynamics of diffusion, transport, or other molecular movements. In contrast to FRAP, the decrease in fluorescence intensity within a defined area is measured.

FCS (Fluorescence Correlation Spectroscopy) FCS allows the analysis of diffusion constants and interactions between different molecules in solution. These spectroscopic techniques register the Brownian random motion of a defined volume *in vitro* with the aid of a focused laser beam.

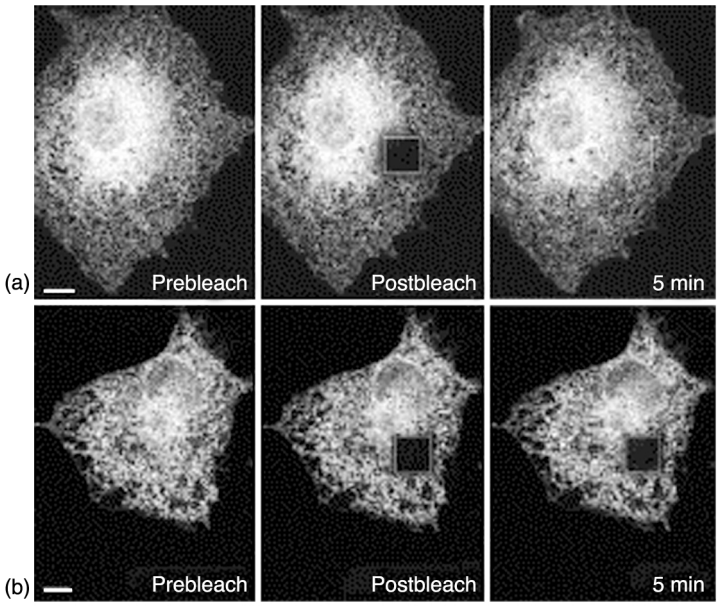


Figure 8.13 FRAP (fluorescence recovery after photobleaching). As an example of a FRAP experiment, the cells shown here have been transfected with a protein GFP construct. To retain proteins in the area of the endoplasmatic reticulum, the cells were cultivated at 40 °C during the experiment. (a) A control cell; (b) cells additionally treated with tunicamycin. The fluorescence recovery after photobleaching shows that the GFP labeled protein is highly mobile in the ER membranes. This is not the case in the absence of tunicamycin. Source: Nehls, S. *et al*. (2000) Dynamics and retention of misfolded proteins in native ER membranes. *Nat. Cell Biol.*, **2**, 288–295. With permission Copyright © 2000, Rights Managed by Nature Publishing Group.

The information is collected from fluctuations in the fluorescence intensity that result from the binding of molecules. The technique was developed in the 1970s, to quantify the diffusion and binding of ethidium bromide to double-stranded DNA. By the combination of FCS with confocal systems (FCM, fluorescence correlation microscopy), the detection volume can be reduced to femtoliters. Besides basic research, FCS/FCM is currently primarily used in drug screening in the medical and pharmaceutical fields.

RICS (Raster Image Correlation Spectroscopy) RICS is a further development of FCS/FCM for the measurement of molecular dynamics and concentrations in living cells. The improvements in sensitivity and speed of confocal systems allows the combination of spectroscopy with methods of live cell imaging. In this way the aggregate state of molecules can be determined with image analysis and special algorithms. Fluctuations in the intensity of fluorescently labeled molecules in cells result, for example, from changes in the conformation of proteins or the binding of proteins to immobilized macrostructures or compartments.

FSM (Fluorescent-Speckle Microscopy) FSM was developed to visualize dynamic processes of polymers, such as the binding and separation of macromolecules *in vitro* and *in vivo* by means of live cell imaging. In this method, derived from the analogous cytometry, fluorescently marked molecules are introduced into a cell by means of microinjection, for example, and incorporated into macromolecular structures. A low proportion of marked molecules compared to the number of unmarked molecules, under 1%, is chosen so that their incorporation in polymers creates a discontinuous labeling pattern. The analysis of time courses provides kinetic information about dynamic processes of polymers (e.g., kinetics of F-actin and microtubules, Figure 8.14) in high spatial and temporal resolution. The full potential of the enormous quantitative potential of this microscopic technique is limited by the demanding mathematical analysis of the localization and photometric characteristics of hundreds of thousands of small emission signals, called speckles. A further difficulty is the conversion of the measured data into biologically relevant information. As a result of the weak fluorescent signal of the speckles, all measurement data from FSM must be calculated using a statistical analysis that employs complicated algorithms. This quantification of a huge mass of data is only possible with robust and automated software solutions. These must be able to extract an enormous number of speckles from time series and analyze their movement. This is difficult because the speckles vary greatly in their fluorescence intensity and the fundamental molecular processes move very quickly.

FRET (Förster Resonance Energy Transfer) FRET is one of the newer methods for the detection of binding and interaction between proteins, lipids, enzymes, DNA, and RNA in living cells. Two molecules of interest are each labeled with fluorescent dyes such that the emission

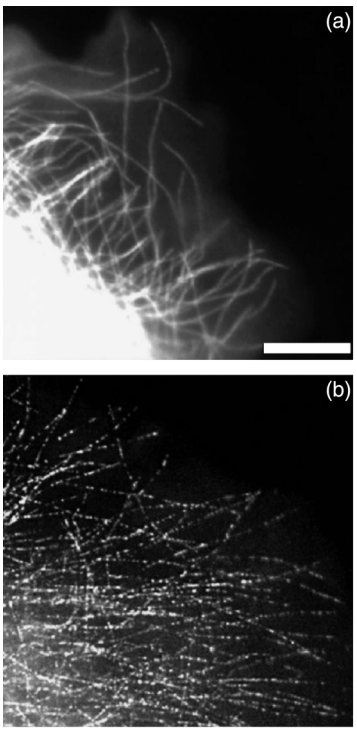


Figure 8.14 Fluorescent-speckle microscopy (FSM). Comparison of conventional epifluorescence microscopy (a) and fluorescence-speckle microscopy (b) using the example of the detection of X-rhodamine-labeled tubulin in the lamellar region of epithelial cells. The labeled tubulin was introduced via microinjection, whereby in (a) 10% of the injected tubulin and in (b) 0.25% of the tubulin injected into each cell was fluorescently labeled. Scale bar 10 µm. Source: Waterman-Storer, C.M. (1998) *Mol. Biol. Cell*, **9**, 3263–3271. With permission, Copyright © 1998 American Society for Cell Biology.

wavelength of one of the dyes, the donor, overlaps with the excitation wavelength of the second dye, the acceptor. When the distance between them is sufficiently small, typically under 10 nm, the donor transfers its energy to the acceptor without emitting a photon. The acceptor emits light, which is detected. Using FRET the relative proximity of the molecules can be detected, well below the optical limits of light microscopy. Possibilities include, for example:

- molecular interactions between two protein partners,
- structural changes in a molecule (such as enzyme activity or DNA/RNA conformation),
- ion concentrations.

FRET, Section 16.7

Spectral Unmixing This is a relatively new method, which allows the use of unusual combinations of fluorescent dyes whose emissions spectra are close to one another, or even overlap. Currently a number of commercial systems are on the market, which detect the spectral composition of the fluorescent light in every detected point of a preparation. In particular, strongly overlapping fluorescence emissions, such as those of the fluorescing proteins CFP, GFP and YFP, can be precisely and efficiently sorted into separate channels of a picture with the help of digital algorithms. In co-localization studies, the artifacts resulting from overlapping absorption and emissions spectra, so-called crosstalk, can be minimized. In FRET experiments, for example using CCP and YFP, the presence of an energy transfer becomes apparent.

Single molecule detection allows the visualization of fluorescently labeled single molecules (e.g., proteins or lipids) and thereby creates the possibility of studying any cellular process on the level of the individual molecules which are central to the process (Figure 8.15a). In this procedure, also referred to as single-dye tracing, the dynamics of individual fluorescently label molecules are detected in 4D. The problem of sterically inhibiting the labeled molecules is significantly reduced by selecting the smallest fluorophore suitable for the experiment. The result is higher resolution that from conventional labeling methods, primarily rely on signal amplification (e.g., indirect immunofluorescence), which frequently masks the details. While the spatial resolution remains limited by the limits of resolution of light microscopes, the precision at which an individual fluorescently labeled molecule can be spatially localized can reach 10 nm or less. The high temporal resolution is achieved by short excitation times. Imaging depends on the detection of the emitted signals per pixel (counts/pixel) by a CCD camera. With the aid of suitable software, the signals can be translated into colors or height maps (Figure 8.15b and c). Time courses allow the analysis of the trajectory of individual molecules (Figure 8.15c and d).

FLIM (Fluorescence Lifetime Imaging) Normally the detection of a fluorescent signal is based on the quantification of the number and spectra of photons emitted by an excited fluorophore. In FLIM, the duration of the excited state is measured. This procedure offers the possibility of detecting different fluorophores, for example, variants of GFP, that aside from overlapping spectral characteristics differ in the duration of their fluorescence. FLIM is therefore well suited

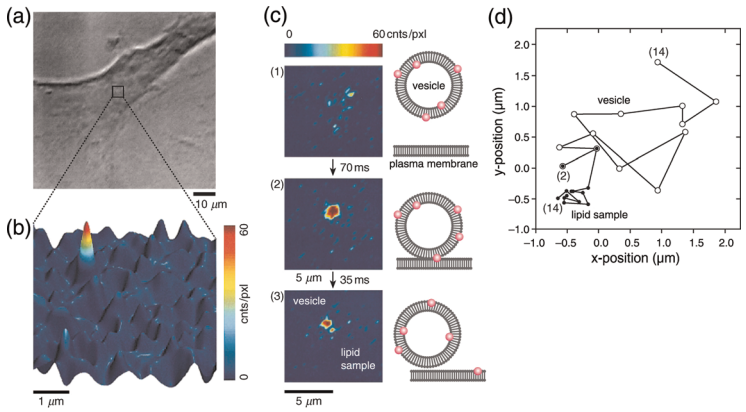


Figure 8.15 Single lipid imaging in cell membranes. As an example of single molecule tracing, human HASM-cells (human smooth muscle cells of the respiratory tract) were pre-treated with Cy5-labeled lipids. (a) The square in the white light image indicates the plasma membrane area selected for fluorescence imaging. (b) The fluorescence image displays a clearly resolved fluorescence peak of a single Cy5-labeled lipid molecule. Very short exposure times (in the range of milliseconds) allow the visualization of single molecule dynamics (c) and subsequently the analysis of individual molecule trajectories (d). Source Schütz, G.J. *et al.* (2000) *EMBO J.*, **19**, 862–901. With permission Copyright © 2000, Rights Managed by Nature Publishing Group

for the measurement of FRET because the duration of the excited state is significantly shorter when FRET events take place. This technique allowed the imaging of the activity of a kinase several years ago. However, various difficulties limit the use of this new technique. Besides lower resolution, FLIM is difficult to use on living cells. Although upgrades for confocal systems are offered commercially, FLIM is a technically very demanding method that requires the use of complicated algorithms for analysis.

Ca^{2+} Imaging The analysis and quantification of, for example, signal transduction processes, is possible indirectly through the detection of changes in fluorescence intensity. This takes advantage of the effects of ions on the fluorescent properties of certain fluorophores, for example, calcium (FURA-2, Fluo-3, 4), or of changes in pH (BCECF). Regardless of the type of indicator used for the imaging procedure, cells are typically loaded with an indicator or are expressing it directly. The read out is captured with a fluorescence microscope. Images are analyzed by measuring fluorescence intensity changes at a single wavelength, or by ratiometric analysis of the fluorescence emission at two wavelengths (ratiometric indicators). The derived fluorescence intensities and ratios are plotted against calibrated values for known Ca^{2+} levels to extrapolate Ca^{2+} concentrations.

The field of live cell imaging is also of great interest to pharmaceutical and biotechnology companies. Many are now using high-throughput and high-content screening platforms for automated analysis of intracellular localization and dynamics. This is paralleled with the increasing development of fluorescent biosensor assays that provide an optical readout of a physiological effect, often based on GFP technology or bioluminescence. Clearly, future developments in this field will be of great interest and benefit to both biotechnology and curiosity-driven research.

Further Reading

Abbe, E. (1873) Beiträge zur theorie des mikroskops und der mikroskopischen wahrnehmung. *Archiv Mikroskop. Anatom.*, **9** (1), 413–468.

Betzig, E. *et al*. (2006) Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, **313** (5763), 1642–1645.

Dunn, G.A. and Jones G.E. (2004) Cell motility under the microscope: vorsprung durch technik. *Nat. Rev. Mol. Cell. Biol.*, **5** (8), 667–672.

Fernández-Suárez, M. and Ting, A.Y. (2008) Fluorescent probes for super-resolution imaging in living cells. *Nat. Rev. Mol. Cell. Biol.*, **9**, 929–943.

Fürstenberg, A. and Heilemann, M. (2013) Single-molecule localization microscopy – near-molecular spatial resolution in light microscopy with photoswitchable fluorophores. *Phys. Chem. Chem. Phys.*, **15** (36), 14919–14930.

Galbraith, C.G. and Galbraith, J.A. (2011) Super-resolution microscopy at a glance. *J. Cell Sci.*, **124** (10), 1607–1611.

Gerlich, D. and Ellenberg, J. (2003) 4D imaging to assay complex dynamics in live specimens. *Nat. Cell. Biol.*, **Suppl**. S14–9.

Goldman, R.D. and Spector, R.D. (eds) (2004) *Live Cell Imaging: A Laboratory Manual*, Cold Spring Harbor Laboratory Press.

Gustafsson, M.G.L. (2005) Nonlinear structured-illumination microscopy: wide-field fluorescence imaging with theoretically unlimited resolution. *Proc. Natl. Acad. Sci. USA*, **102** (37), 13081–13086.

Hell, S.W. (2007) Far-field optical nanoscopy. *Science*, **316** (5828), 721–723.

Hell, S.W. and Wichmann, J. (1994) Breaking the diffraction resolution limit by stimulated emission: stimulated-emission depletion fluorescence microscopy. *Opt. Lett.*, **24** (14), 780–782.

Jacobs, R.E. *et al*. (2003) MRI: volumetric imaging for vital imaging and atlas construction. *Nat. Rev. Mol. Cell. Biol.*, Suppl. SS10–6.

Krichevsky, O. and Bonnet, G. (2002) Fluorescence correlation spectroscopy: the technique and its applications. *Rep. Prog. Phys.*, **65**, 251–297.

Lippincott-Schwartz, J. *et al*. (2003) Photobleaching and photoactivation: following protein dynamics in living cells. *Nat. Cell. Biol.*, Suppl. S7–14.

Lippincott-Schwartz, J. and Patterson, G.H. (2003) Development and use of fluorescent protein markers in living cells. *Science*, **300** (5616), 87–91.

Lippincott-Schwartz J. *et al*. (2001) Studying protein dynamics in living cells. *Nat. Rev. Mol. Cell. Biol.*, **2** (6), 444–456.

Miyawaki, A. *et al*. (2003) Lighting up cells: labelling proteins with fluorophores. *Nat. Cell. Biol.* Suppl. S1–7.

Phair, R.D. and Misteli, T. (2001) Kinetic modelling approaches to in vivo imaging. *Nat. Rev. Mol. Cell. Biol.*, **2** (12), 898–907.

Roy, R., Hohng, S., and Ha, T. (2008) A practical guide to single-molecule FRET. *Nat. Methods*, **5**, 507–516.

Rust, M.J., Bates, M., and Zhuang, X. (2006) Sub-diffraction limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods*, **3**, 793–795.

Sako, Y. and Yanagida, T. (2003) Single-molecule visualization in cell biology. *Nat. Rev. Mol. Cell. Biol.*, Suppl. SS1–5.

Schütz, G.J. and Schindler, H. (2002) Single dye tracing for ultrasensitive microscopy on living cells, single molecule detection in solution, in *Single Molecule Detection in Solution* (eds Ch. Zander, J. Enderlein, and R.A. Keller), Wiley-VCH Verlag GmbH, Weinheim.

Stelzer, E.H. (2014) Better imaging through chemistry. *Cell*, **159** (6), 1243–1246.

Steyer, J.A. and Almers, W. (2001) A real-time view of life within 100 nm of the plasma membrane. *Nat. Rev. Mol. Cell. Biol.*, **2** (4), 268–275.

Tsien, R.Y. (1998) The green fluorescent protein. *Annu. Rev. Biochem.*, **67**, 509–544.

Tsien, R.Y. (2003) Imagining imaging's future. *Nat. Rev. Mol. Cell. Biol.*, Suppl. SS16–21.

Wiegräbe, W. (2000) Fluorescence correlation microscopy: probing molecular interactions inside living cells. *Am. Lab.*, **32** (18), 44–47.

Cleavage of Proteins

9

Josef Kellermann

Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

The cleavage of proteins, or proteolysis, is the breakdown of proteins or peptides into smaller fragments or even single amino acids. In general the peptide bond is hydrolyzed by enzymes, so-called proteases, as well as by non-enzymatic, chemical methods.

In vivo proteolysis has a variety of different tasks. In the digestive organs several enzymes, like pepsin in the stomach, trypsin in the intestine, or chymotrypsin in the pancreas, break down proteins from food, making the amino acids available for metabolic turnover. Proteolysis is important during the regulation of physiological and cellular processes and helps to remove unnecessary or abnormal proteins as a kind of cellular maintenance such as occurs in the proteasome. Proteolytic processing is also a frequently used way to activate precursor proteins from zymogens.

In vitro proteolysis is used for research purposes in the analytical laboratory. Protein tags are removed from fusion proteins after their purification. Undesired enzymatic activity or unwanted proteins can be removed during the preparation of DNA or RNA. Proteins can be partially inactivated, such as happens when DNA polymerase I is treated with subtilisin to yield the Klenow fragment, which retains its polymerase activity but lacks its 5′-exonuclease function. The stability of folded domains is tested under a wide range of conditions. During crystallization processes topological studies help to increase the success rate of the experiments.

Above all, proteolysis has for many years played an important role during the structure analysis of proteins. To identify or to confirm the primary structure of proteins with Edman degradation or mass spectrometry it is necessary to break down large proteins into peptides that are suitable for analysis by particular methods. To identify each single amino acid in a protein and to arrange the analyzed fragments in the right sequence, overlapping peptide fragments have to be generated with several different enzymes (Figure 9.1).

Protein Sequence Analysis, Chapter 14

9.1 Proteolytic Enzymes

Cleaving peptide bonds of proteins with high specificity makes proteolytic enzymes important tools in the structure elucidation of proteins. According to the extent of the cleavage, complete and incomplete (limited) proteolysis can be distinguished. A proteolytic reaction, where hydrolysis of a specific peptide bond reaches 100%, creates an equimolar set of peptides that is highly characteristic for a specific protein. Limited proteolysis does not reach its endpoint and results in different peptide patterns depending on the amount of protease, the addition of an inhibitor, or changes in reaction parameters like pH or temperature. Limited proteolysis is important in order to study early cleavage products or easy accessible cleavage sites of a protein, or to generate large peptide fragments. Limited proteolysis experiments can provide accurate definition of domain boundaries during the structure elucidation of proteins and help in studying the topology of native proteins. The proteolytic accessibility of native proteins is

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

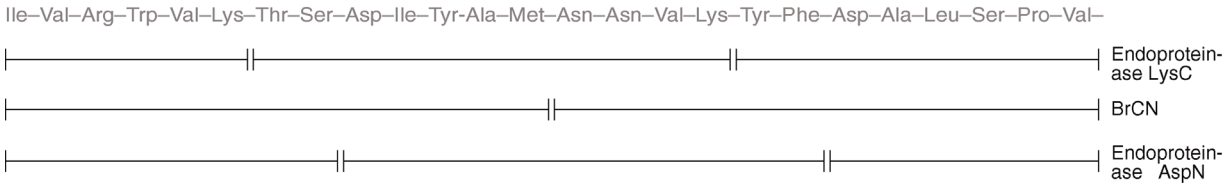


Figure 9.1 Fragmentation of proteins: cleavage of a protein with three different enzymes to generate overlapping fragments.

often limited owing to a compact higher order structure. There are many examples of this kind of limited proteolysis *in vivo*, such as zymogen activation, prohormone processing, or cleaving out biological active peptides from a precursor protein. During primary structure analysis and to create peptide maps proteins have to be fragmented completely, to generate an exactly defined and reproducible peptide pattern.

9.2 Strategy

Protein Purification, Chapter 1

Most of the technologies used for protein characterization need proteins that are almost purified to homogeneity. Starting with complex protein mixtures, such as present in a complete cell, several chromatographic or electrophoretic purification steps are necessary until a protein is pure enough for fragmentation and further characterization. Depending on the purification strategy proteins have to be cleaved in solution, in a polyacrylamide gel matrix, or blotted on a membrane (Figure 9.2)

Cleavage in Solution Proteolysis is commonly carried out in solution when a protein is already purified to homogeneity. Often the problem is the solubility of a protein. The solubilization buffer therefore often contains detergents and salts that may affect either protease activity or a subsequent chromatographic separation.

Electrophoretic Techniques, Chapter 11

Cleavage on Membranes Proteins can be transferred onto inert membranes after separation by polyacrylamide gel electrophoresis or 2D gel electrophoresis followed by staining with Coomassie, Amido black, Ponceau red, or any immune staining procedure. The transfer on a membrane allows for high flexibility of the subsequent methods used. A protein can be N-terminal sequenced by Edman degradation or a N-terminal blocked amino acid can be enzymatically removed, followed by sequencing. Just as well, proteins can be enzymatically or chemically cleaved on the membrane and the resulting peptides can be eluted and further analyzed by MS or LC-MS methods.

Cleavage in SDS-Polyacrylamide Gels Electrophoretic separation by polyacrylamide gel electrophoresis or 2D gel electrophoresis nowadays is the most powerful and effective strategy to fractionate or purify proteins on a micro scale. Following separation proteins usually are stained

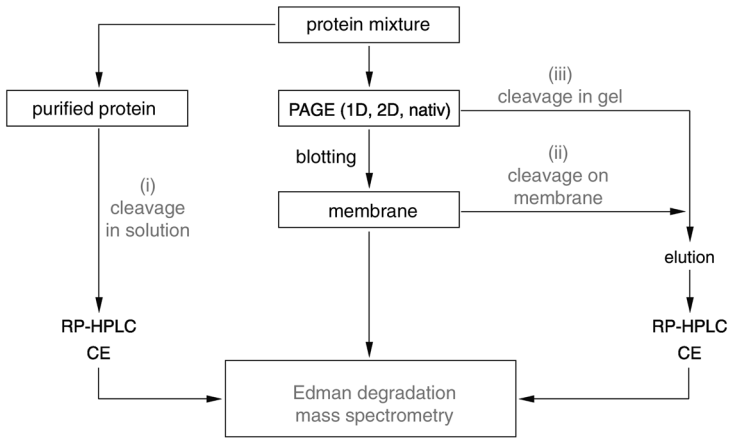


Figure 9.2 Strategy of protein characterization. Fragmentation is dependent on the method used for purification: (i) cleavage in solution; (ii) cleavage on a membrane; (iii) cleavage in polyacrylamide matrix.

with Coomassie or digestion compatible silver stains. Gel slices or single spots are then cut into small pieces, washed with buffer, and shrunk either with organic solvents like acetonitrile or by drying. Subsequently, pieces are rehydrated in buffer, containing the required enzyme and incubated at appropriate conditions, in good accordance with the optimal conditions for the used enzyme. Peptides then are eluted from the gel matrix and can be further analyzed by MS or LC-MS methods.

Mass Spectrometry, Chapter 15

Chromatographic Separation Techniques, Chapter 10

9.3 Denaturation of Proteins

Compact secondary and tertiary structures of native proteins often result in poor cleavage yields since cleavage sites are not accessible for proteases. Denaturation using detergents and urea or guanidine hydrochloride destroys higher order structures, which is often accompanied by the loss of activity of the protein. Often, several detergents are used during purification of membrane proteins. Urea might be contaminated with cyanate ions, which can block amino groups by carbamylation. Therefore, a solution of 6 M guanidine hydrochloride is preferred as denaturing agent. This solution diluted to 1 M is for many proteases compatible with their activity. The most frequent reason why a protein is resistant to enzymatic degradation is based on disulfide bonds, which stabilize the three-dimensional structure of the native peptide chain. To make proteins accessible for proteases it is essential to cleave their disulfide bridges irreversibly.

9.4 Cleavage of Disulfide Bonds and Alkylation

Disulfide bridges are cleaved following denaturation (Figure 9.3 and Table 9.1). This can be achieved either by oxidation (Figure 9.3a), where disulfides are oxidized to cystic acid by performic acid, or by reduction with reducing agents such as dithiothreitol (DTT), 2-mercaptoethanol, or tributylphosphine (Figure 9.3b). Because of its volatile properties, tributylphosphine does not have to be removed before further treatment of a protein. To avoid uncontrolled rearrangement of disulfides the free SH-groups have to be stabilized by alkylation (Figure 9.4).

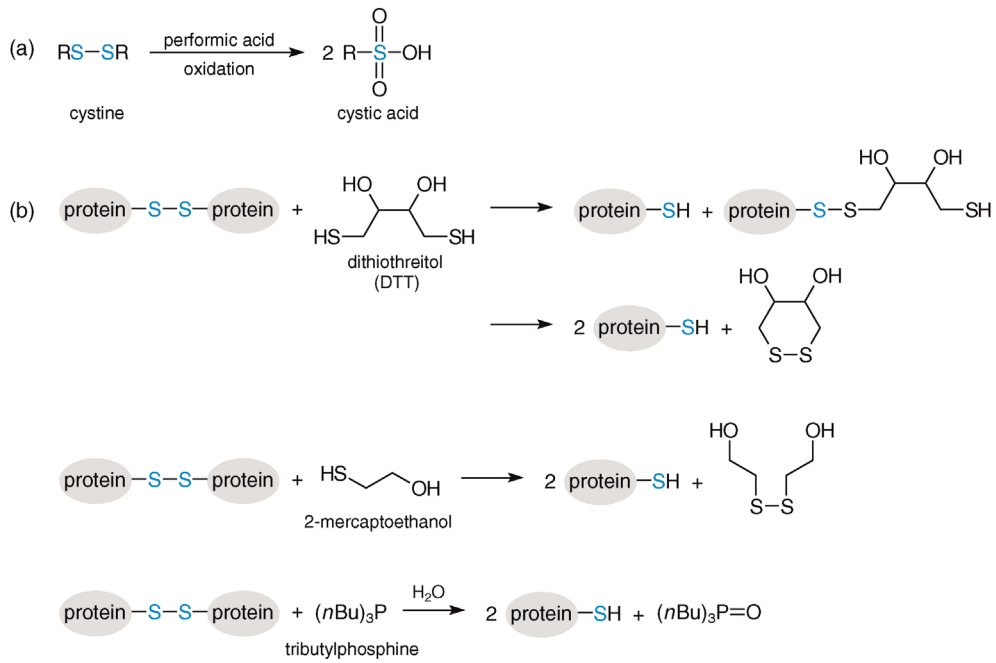


Figure 9.3 Cleavage of disulfide bonds: (a) oxidation of disulfide bonds; reaction of cysteine with performic acid to cystic acid; (b) reduction of disulfide bonds: reaction with dithiothreitol, 2-mercaptoethanol, and tributylphosphine (R = peptide chain).

Table 9.1 Chemical modification of cysteine residues.

Reagents	Modified cysteine residue
Performic acid	Cysteic acid
Sulfite	*S*-Sulfocysteine
Iodoacetic acid	*S*-(Carboxymethyl)cysteine
Iodoacetamide	*S*-(Carboxamidomethyl)cysteine
Ethyleneimine	*S*-(2-Aminoethyl)ethylcysteine
4-Vinylpyridine	4-(Pyridylethyl)cysteine
Acrylamide	Cys-*S*-β-propionamide

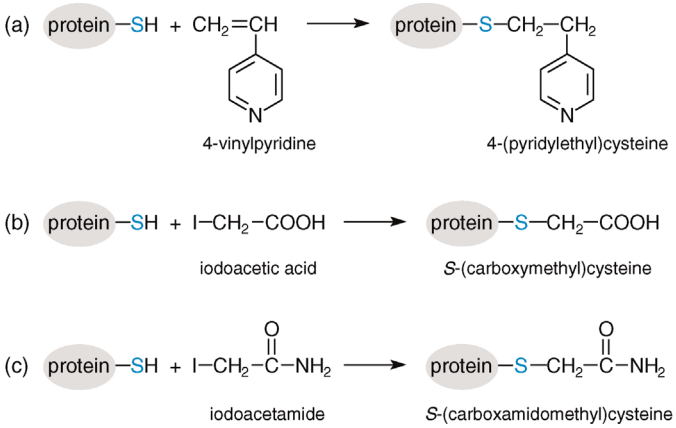


Figure 9.4 Alkylation of cysteine residues: reaction of cysteine and (a) 4-vinylpyridine to yield 4-pyridylethyl cysteine, (b) iodoacetic acid to yield *S*-(carboxymethyl)-cysteine, and (c) iodoacetamide to yield *S*-(carboxamidomethyl)-cysteine.

Immediately after the reaction of 4-vinylpyridine excess reagent has to be removed from the reaction mixture, otherwise side reactions with His, Trp and Met might happen. Peptides, which have reacted with 4-vinylpyridine show an absorption maximum at 256 nm which can be used to selectively isolate these peptides.

This is also important for Edman degradation, since unmodified cysteine residues cannot be identified during sequence analysis. The mainly used alkylation reagents are 4-vinylpyridine, iodoacetic acid, and 2-iodoacetamide. During electrophoresis free acrylamide monomers can react with thiols of a protein. Therefore, cysteine residues should be alkylated before electrophoresis if mass spectrometry or Edman sequencing is the final analysis tool. Many additional methods are described elsewhere, none of which play a major role in protein analysis.

9.5 Enzymatic Fragmentation

Protein Sequence Analysis, Chapter 14

Mass Spectrometry, Chapter 15

If no information on a protein is available or when a complete proteome has to be cleaved, the average frequency of an amino acid in a protein can help to decide which enzyme can generate suitable fragments. Few and long fragments are generated when enzymes are used to cleave very specific or rare amino acids. Many short fragment are produced when enzymes with low specificity are used to cleave many different amino acids. Table 9.2 illustrates the average frequency of amino acid residues calculated from the NBRF-PIR database. A hypothetical protein, containing 300 amino acids, is cleaved with the most frequent enzymes, resulting in a specific number of fragments of particular length.

The cleavage behavior of a protein towards an enzyme is characteristic and highly reproducible under defined conditions. The separation of the resulting fragments using SDS-PAGE, capillary electrophoresis, or chromatographic methods (HPLC) results in a characteristic peptide pattern (finger prints or peptide maps, Figure 9.5) that can be analyzed by Edman sequencing or mass spectrometry.

9.5.1 Proteases

According to their mode of action and their active center proteases are classified as *serine-, cysteine-, aspartic acid-* and *metallo-proteases*. Depending on their cleavage site, *endo-* and *exo-proteinases* can be distinguished.

Table 9.2 Theoretical number and length of peptide fragments of a hypothetical protein with 300 amino acid residues, cleaved by a particular enzyme or reagent, calculated according to NBRF-PIR.[a)]

Enzyme or reagent	Cleaves specifically at	Average fragment length	Number of fragments
Chymotrypsin	Leu, Phe, Trp, Tyr	6	54
Trypsin	Lys, Arg	9	35
Endoproteinase Glu-C	Glu	15	20
Endoproteinase Lys-C	Lys	16	19
Endoproteinase Arg-C	Arg	18	17
Endoproteinase Asp-N[b)]	Asp	18	17
Cyanogen bromide	Met	38	8
BNPS-skatole	Trp	60	5

a) National Biomedical Research Foundation – Protein Identification Resource.
b) Proteolysis N-terminal of Asp.

Endoproteases cleave the protein backbone at internal amino acid residues, which are specific for each protease, generating a highly specific peptide pattern for each protein. They are, therefore, preferably used during primary structure analysis. Many of these proteases cleave C- or N-terminal of charged amino acid residues (endoproteinase Lys-C, trypsin, endopeptidase Asp-N). Proteases with less specificity like chymotrypsin, thermolysin, or pepsin cleave proteins at several different amino acids, yielding more and shorter peptide fragments (Table 9.3).

Endoproteases

Chymotrypsin Chymotrypsin (25 kDa) is a serine protease that hydrolyzes peptide bonds at the C-terminal of Tyr, Phe, and Trp. Leu, Met, Ala, Asp, and Glu can also be hydrolyzed but with less efficiency.

Elastase Elastase (25 kDa), which is also a serine protease, hydrolyzes peptide bonds at the C-terminal of uncharged aromatic amino acids (Ala, Val, Ile, Leu, Gly, and Ser). Elastase is often used to solubilize membrane proteins.

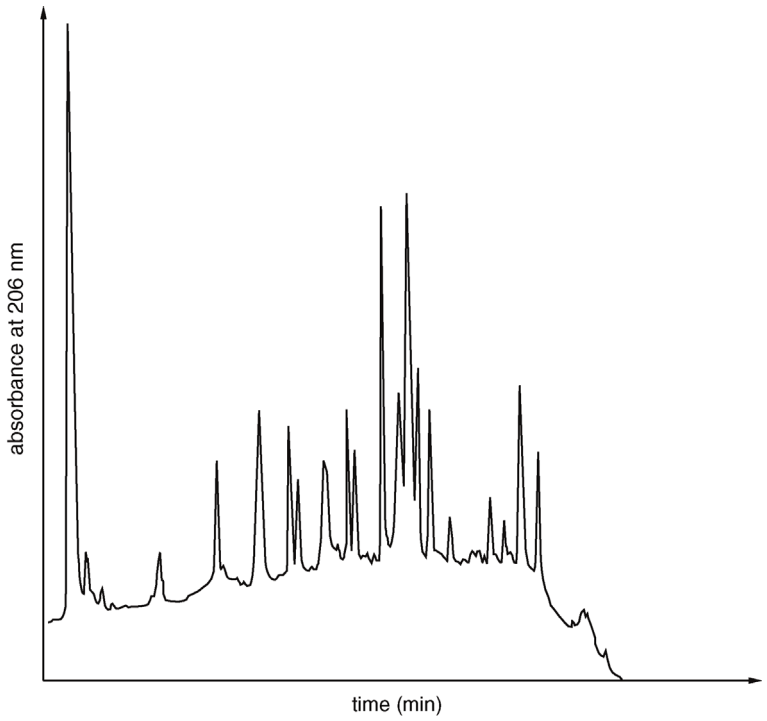


Figure 9.5 Fingerprint of a protein cleaved with endoproteinase Lys-C (40 kDa). Reversed phase chromatography of 30 pmol of protein (Superspher® 60 RP select B (Merck), 2 × 125 mm) buffer A: 0.1% trifluoroacetic acid (TFA); buffer B: 0.85% TFA in acetonitrile; gradient 1% min^{-1}, flow rate 0.3 ml min^{-1}.

Table 9.3 Enzymes used in protein structure analysis.

Enzyme	EC number	Type	Specificity	pH optimum	Inhibitors[a)]
Endopeptidases					
Chymotrypsin	3.4.21.1	Serine	Tyr, Phe, Trp	7.5–8.5	Aprotinin, DFP, PMSF
Trypsin	3.4.21.4	Serine	Arg, Lys	8.0–9.0	TLCK, DFP, PMSF,
Endoproteinase Glu-C	3.4.21.19	Serine	Glu	8.0	DFP, α2-macroglobulin, 3,4-dichloroisocoumarin
Endoproteinase Lys-C	3.4.21.50	Serine	Lys	7.5–8.5	DFP, TLCK, aprotinin, leupeptin
Endoproteinase Arg-C	3.4.22.8	Cysteine	Arg	8.0–8.5	Oxidation reagent EDTA, Co^{2+}, Cu^{2+}, citric acid, borate
Endoproteinase Asp-N		Metallo	Asp[b)], Cystic acid	6.0–8.0	EDTA, *o*-phenanthroline
Elastase	3.4.21.36	Serine	Ala, Val, Ile, Leu, Gly	8.9	DFP, α1-antitrypsin, PMSF, elastinal
Pepsin	3.4.23.1	Aspartate	Phe, Met, Leu, Trp	2.0–4.0	Pepstatin, 4-bromphenacyl bromide
Subtilisin	3.4.21.14	Serine	Almost all AA	7.0–11.0	DFP, PMSF, indole, phenol
Thermolysin	3.4.24.4	Zn-Metallo	Hydrophobic AA	7.0–9.0	Chelate formers (EDTA), phosphoamidon
Elastase	3.4.21.36	Serine	Uncharged, non-aromatic AA	8.8	DFP, α1-antitrypsin, PMSF
Papain	3.4.22.2	Cysteine	Arg, Lys, Glu, His, Tyr	7.0–9.0	Iodoacetic acid, iodoacetamide, TPCK, TLCK,
Pronase		mixture	All AA	7.5	No special inhibitors
Proteinase K	3.4.21.14	Serine	hydrophobic, aromatic AA	7.0	SH-blocker (iodoacetamide)
Thrombin	3.4.21.5	Serine	Arg	7.5	DFP, TLCK, PMSF, leupeptin, STI
Factor X	3.4.21.6	Serine	Ile-Glu-Gly-Arg	8.3	DFP, PMSF, STI
Enterokinase	3.4.21.9	Serine	$(Asp)_4$-Lys	8.0	DFP, TLCK
Exopeptidases					
C-terminal					
Carboxypeptidase P	3.4.16.1	Serine	PrO-Xaa-COOH	4.0–5.5	DFP, iodoacetic acid
Carboxypeptidase C	3.4.16.1	Serine	C-terminal peptide unspecific	4.0–5.0	
Carboxypeptidase Y	3.4.16.1	Serine	C-terminal peptide unspecific	5.5–6.5	DFP, PMSF, ZPCK, aprotinin
Carboxypeptidase A	3.4.17.1	Zn-Metallo	C-terminal peptide unspecific	7.0–8.0	Chelate former (pyrophosphate, oxalate)
Carboxypeptidase B	3.4.17.2	Zn-Metallo	Basic AA, C-terminal	7.0–9.0	Chelate former, basic AA
N-terminal					
Acylamino acid-releasing enzyme	3.4.19.1	Serine	*N*-Acyl-AS	7.5–9.0	DFP
Pyroglutamate aminopeptidase	3.4.19.3	Cysteine	Pyroglutamate	7.0–9.0	SH-blocker (iodoacetamide)
Cathepsin C	3.4.14.1	Cysteine	N-terminal dipeptides	5.5	Iodoacetic acid, formaldehyde
Glucosidases					
N-Glycosidase A and F	3.5.1.52		*N*-Acetyl-β-D-glucosamine	4.5–7.0	
O-Glycosidase	3.2.1.97		D-Galactosyl-*N*-acetyl-D-galactosamine	6.0	
Phosphatases					
Acidic Phosphatase	3.1.3.2		Ortho-phosphomonoester	3.0–6.0	Fluoride, molybdate, orthophosphate
Alkaline Phosphatase	3.1.3.1		Ortho-phosphomonoester	7.0	

a) DFP = diisopropyl fluorophosphate, TLCK = L-1-chloro-3-(4-tosylamido)-7-amino-2-heptanone hydrochloride, PMSF = phenylmethylsulfonyl fluoride, TPCK = L-1-chloro-3-(4-tosylamido)-4-phenyl-2-butanone, ZPCK = carbobenzoxy-L-phenylalanine-chloromethyl ketone.

b) Cleaves N-terminal of Asp.

Endoproteinase Arg-C Endoproteinase Arg-C (30 kDa) is a serine proteinase with high specificity, cleaving only the peptide bonds C-terminal of arginine. Because of its high specificity and resistance against autoproteolysis Arg-C is one of the most important proteases in primary structure analysis.

Endoproteinase Asp-N This metalloproteinase (27 kDa) hydrolyzes peptide bonds N-terminal of aspartic acid and cystic acid.

Endoproteinase Glu-C Endoproteinase Glu-C, also called protease V8 (27 kDa), is a serine protease that hydrolyzes peptide bonds C-terminal of glutamic acid either in ammonium hydrogen carbonate (pH 7.8) or in ammonium acetate (pH 4). By using phosphate buffer (pH 7.8), aspartic acid is cleaved, too, at its C-terminus.

Endoproteinase Lys-C This serine proteinase hydrolyzes specific amide-, ester-, and peptide-bonds C-terminal of lysine. Additional crosslinking stabilizes this enzyme and protects it against autoproteolysis, making this protease a valuable tool for microanalytical methods.

LysargiNase LysargiNase, also described as ulilysin, is a metalloproteinase isolated from *Methanosarcina acetivorans* that cleaves proteins before lysine and arginine. LysargiNase generated peptides have N-terminal lysine or arginine residues and fragment with b-ion-dominated spectra. This improves protein C-terminal identification. LysargiNase also cleaves at methylated and dimethylated lysine and arginine residues.

Papain Papain is a cysteine proteinase that cleaves peptides C-terminal of Arg, Lys, Glu, His, Gly, and Tyr. However, at longer incubation times nearly all peptide bonds are cleaved. Therefore, papain is often used for the total hydrolysis of proteins. Additionally, papain shows esterase and transamidase activity. Activity can be inhibited by SH-reactive agents. Papain is often used to specifically cleave native immunoglobulins into Fab and Fc fragments. As well as elastase, papain is used to solubilize membrane proteins.

Pepsin Pepsin is an aspartic acid protease with broad specificity; it cleaves the peptide bonds of Phe, Met, Leu, or Trp to other hydrophobic amino acids. Its pH optimum of 2.0 is remarkable. Pepsin is one of the few enzymes with an acidic pH optimum. Similarly to papain, pepsin liberates active Fab_2 fragments.

Pronase Pronase is an unspecific mixture of different proteases from *Streptomyces griseus*. Pronase is used especially for the total hydrolysis of proteins. Together with trypsin or collagenase, pronase is used to dissociate tissue. There is no known inhibitor for this enzyme.

Subtilisin Subtilisin is a serine proteinase that is mainly used for total hydrolysis. A remarkable property is the extreme pH range over which this enzyme is active (pH 7–11).

Thermolysin Thermolysin (37 kDa) is a Zn-protease with low specificity. It hydrolyzes especially at amino acid residues with hydrophobic, large side chains like Ile, Leu, Met, Phe, Trp, and Val (except when Pro is on the C-terminal side). One advantage is its extreme thermostability (4–80 °C).

Trypsin Trypsin (23 kDa) catalyzes the cleavage of peptide bonds C-terminal of Arg and Lys as well as the corresponding amides and ester. It is one of the most applied enzymes in protein analysis to create peptide mass finger prints, since a defined charge distribution is achieved with Lys or Arg at the C-terminus of the peptide.

Proteinase K Proteinase K is a protease with low specificity and is often used during the isolation of RNA and DNA to remove and inactivate proteins. Proteinase K degrades proteins at the C-terminal of hydrophobic, aliphatic, and aromatic amino acid residues.

Factor Xa, Thrombin, and Enterokinase A special application of these endoproteinases is during the isolation of recombinant proteins. Cloning of cleavage sites with very high specificity allows the desired protein to be cleaved from a fusion protein after its purification.

Exoproteases Exopeptidases degrade proteins from their amino- or carboxyl-termini. They are used in protein chemistry to make N-terminal blocked proteins accessible to Edman-degradation. Most frequently used is the acylamino acid releasing enzyme, which removes N-terminal acetyl-groups, and pyroglutamic acid aminopeptidase, which removes pyroglutamic acid residues. Often proteins have to be fragmented with endoproteinases before applying these exoproteases making them accesable to hydrolysis.

C-Terminal Sequence Analysis, Section 14.2.3

Carboxypeptidases Carboxypeptidases are used to degrade proteins from their C-terminus to obtain sequence information. Since this is not a sequential reaction, such as during the Edman chemistry, but a continuous one the time course of the enzymatic reaction has to be recorded to calculate the amount of released amino acids by amino acid analysis. Usually, a mixture of carboxypeptidase A, B, and Y is used each component having a different specificity. Carboxypeptidase A slowly degrades Gly, Asp, Glu, Cys, and $CysSO_3H$ at the C-terminus but is not able to cleave off Arg and Pro. Carboxypeptidase B particularly releases Lys and Arg. Carboxypeptidase Y has a broad cleavage spectrum, releasing nearly all amino acids except Gly and Asp. This sequencing method, however, is increasingly being replaced by mass spectrometric methods.

Amino Acid Analysis, Chapter 13

Carbohydrate Analysis, Chapter 23

Glycosidases Carbohydrate side chains often hinder proteases in cleaving the protein backbone and prevent identification of the amino acid residue involved in the protein/carbohydrate linkage during Edman degradation. Glycosidases are a group of enzymes that do not belong to proteases but are indispensable for structure analysis. Especially, N-glucosidases (A and F) are able to completely remove carbohydrate side chains from a protein (Asn), producing aspartic acid and ammonia. O-Glucosidases completely remove galactosamine linked to threonine.

Phosphatases The activation of many proteins is regulated by the phosphorylation of serine, threonine, or tyrosine residues. This modification can be identified by enzymatic removal using phosphatases, resulting in a change of molecular weight, which can be determined by mass spectrometry, and a change of mobility during isoelectric focusing, which is caused be a change of the electric charge.

Classification of Proteolytic Enzymes There are four classes of proteases, which differ in their mode of action. These classes, in turn, are classified in six families having different amino acid arrangements around their active center (Table 9.4). They are named according to the amino acid that is directly involved in the catalytic mechanism (e.g., serine protease). Serine proteases are further subdivided into bacterial and mammalian proteases as they differ in their three-dimensional structure despite having the same active site. The same is true for metalloproteases.

The society for classification of enzymes (the Enzyme Commission, EC) has established a classification scheme with a four-figure number that clearly characterizes any enzyme.

Table 9.4 Classification of proteolytic enzymes.

Family	Enzymes	Active center
Serine protease I	Chymotrypsin, trypsin, elastase, pancreatic kallikrein	Asp^{102}, Ser^{195}, His^{57}
Serine protease II	Subtilisin	Asp^{32}, Ser^{221}, His^{64}
Cysteine protease	Papain, cathepsin	Cys^{25}, His^{159}, Asp^{158}
Aspartic protease	Pepsin, renin	Asp^{33}, Asp^{213}
Metalloprotease I	Carboxypeptidase A	Zn, Glu^{270}, Try^{248}
Metalloprotease II	Thermolysin	Zn, Glu^{143}, His^{231}

The first number describes six main categories, of which hydrolases (EC 3) are the most important for structure analysis of proteins. The second number describes the kind of bond (e.g., peptide bond EC3.4). The third number characterizes the catalytic center of the enzyme (EC 3.4.21 for serine proteases or EC 3.4.24 for metalloproteases). The last number is a kind of serial number within it subclass (EC 3.4.21.4 for trypsin, see also Table 9.3).

9.5.2 Conditions for Proteolysis

The most important parameters of the enzymatic hydrolysis are the buffer, its molarity, pH, temperature, incubation time, and the molar ratio of an enzyme to its substrate. Table 9.5 summarizes some proteolysis conditions for different enzymes. The kind of buffer needed is normally defined by the pH range that is necessary for the catalytic reaction. Most enzymes retain their activity even in the presence of detergents. Organic solvents like acetonitrile or isopropanol improve the solubility of a protein without affecting their activity. Even the presence of reducing agents is not harmful for many enzymes. Incubation times between 4 and 16 h at 37 °C in most cases are sufficient for a complete hydrolysis. The buffer volume should be kept as small as possible to achieve high substrate and enzyme concentrations. If high enzyme to substrate ratios are chosen, enzymes should be applied that are protected against autoproteolysis (e.g., endoproteinase Lys-C and endoproteinase Glu-C). All enzymes should be freshly dissolved immediately before use.

Some enzymes need particular ions for their activity. Trypsin, for example, is only active in the presence of Ca^{2+} ions (2 mM) and pyroglutamic acid aminopeptidase needs thiols for its activity.

To hydrolyze blotted, membrane bound proteins it is necessary to saturate the membrane before the addition of the enzyme with quenching reagents (e.g., polyvinylpyrrolidone (PVP), such as PVP40) to avoid unspecific binding of the enzyme to the membrane. Excess reagent has to be removed by extensive washing since PVP40 would disturb a subsequent chromatography step. Alternatively, hydrogenated Triton X-100 (RTX-100) can be added to the cleavage buffer; it does not absorb in the UV during subsequent chromatography.

Table 9.5 Proteolysis conditions.

Component	Application
Buffer	
0.1 M Ammonium hydrogen carbonate	Suitable for many enzymes at about pH 8
N-Methylmorpholine	
Detergents	
SDS 0.1%	Suitable for most enzymes (for subtilisin and endoprotease Lys-C: up to 1%)
CHAPS, octylglucoside, NP40 up to 2%	
Organic solvents	
Acetonitrile 20%	Endoprotease Glu-C, pepsin, trypsin (up to 40%), endoprotease Lys-C
Isopropanol 20%	Endoprotease Asp-N, subtilisin, thermolysin, papain, elastase
Reducing agents	
Mercaptoethanol 0.5%	Endoprotease Lys-C and Glu-C
Mercaptoethanol 1%	Endoprotease Asp-N
Cleavage time	
4–16 h at 37 °C in most cases is sufficient for complete reaction.	
Enzyme/substrate-ratio	
Cleavage in solution: 1 : 20 to 1 : 100	
Cleavage in gel or on membrane: 1 : 1 to 1 : 10	
(autoproteolysis at high concentrations!)	

Enzymatic in-gel cleavage requires extensive alternate washing of the gel pieces with cleavage buffer and acetonitrile to shrink the gel piece. The shrunken gel draws the cleavage buffer together with the enzyme. The elution of the peptide fragments is done with organic solvent and formic acid or trifluoroacetic acid.

To be able to distinguish peptides, impurities, and autolysis products, a reference sample without protein should be processed as well.

9.6 Chemical Fragmentation

Although proteins or peptides can be hydrolyzed by enzymes at numerous amino acid residues, further sites can be cleaved by chemical methods, having the advantage of being tolerant towards salts and detergents. Although a great number of chemical cleavage methods have been described, only a few of them are suitable for routine applications. Low cleavage yields, low specificity, and undesired side reactions lead to poor reproducibility, making them uninteresting for structure analysis.

Cleavage of Met-X Bond with Cyanogen Bromide The most frequently used reagent is cyanogen bromide. It has the advantage of cleaving almost quantitatively and highly specifically the peptide bond C-terminal of methionine (Met-Xaa). There is no enzyme available that can cleave Met with high specificity. The cleavage results in poor yields when threonine or serine is involved in the peptide bond Xaa. Since methionine is a relatively rare amino acid the size of the produced fragments is above average. The reaction takes place in 70% formic acid, a good solvent for many proteins. As a side reaction the peptide bonds between aspartic acid and proline are cleaved partially by acid hydrolysis. The reagents are volatile and, therefore, well suited for the subsequent purification steps. The selectivity of cyanogen bromide is based on the electrophilic attack of CNBr towards the sulfur of the methionine (Fig. 9.6) forming a sulfonium ion, release of methyl thiocyanate, building an intermediate imino ring, hereby cleaving the peptide bond. A new amino group is formed as well as homoserine as new C-terminal amino acid residue, being in equilibrium with homoserine lactone.

The cleavage usually is performed using between 10- and 100-fold molar excess of cyanogen bromide over methionine in the dark in an oxygen-free environment for 2–16 h. A higher excess of CNBr or longer cleavage time leads to side reactions, especially involving tryptophan residues. The cleavage can also be performed on PVDF-membranes or in polyacrylamide matrices.

Partial Acid Hydrolysis The Asp-Pro peptide bond can also be cleaved with 70% formic acid or trifluoroacetic acid. Since this reaction is not complete, heterogeneous fragments are generated when more than one cleavage site is in one protein.

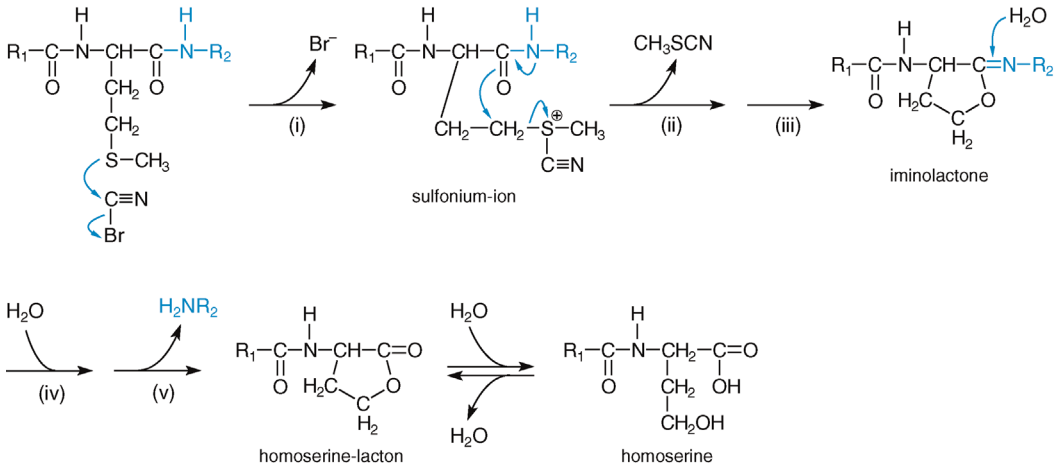


Figure 9.6 Reaction scheme of cleavage with cyanogen bromide. Electrophilic attack (i) of CNBr towards the sulfur of the methionine, forming a sulfonium ion (ii); release of methyl thiocyanate, building an intermediate imino ring (iii); hydrolysis of the iminolactone ring and cleavage of the peptide bond (iv); formation of a new amino group (v); homoserine is formed as new the C-terminal amino acid residue, being in equilibrium with homoserine lactone.

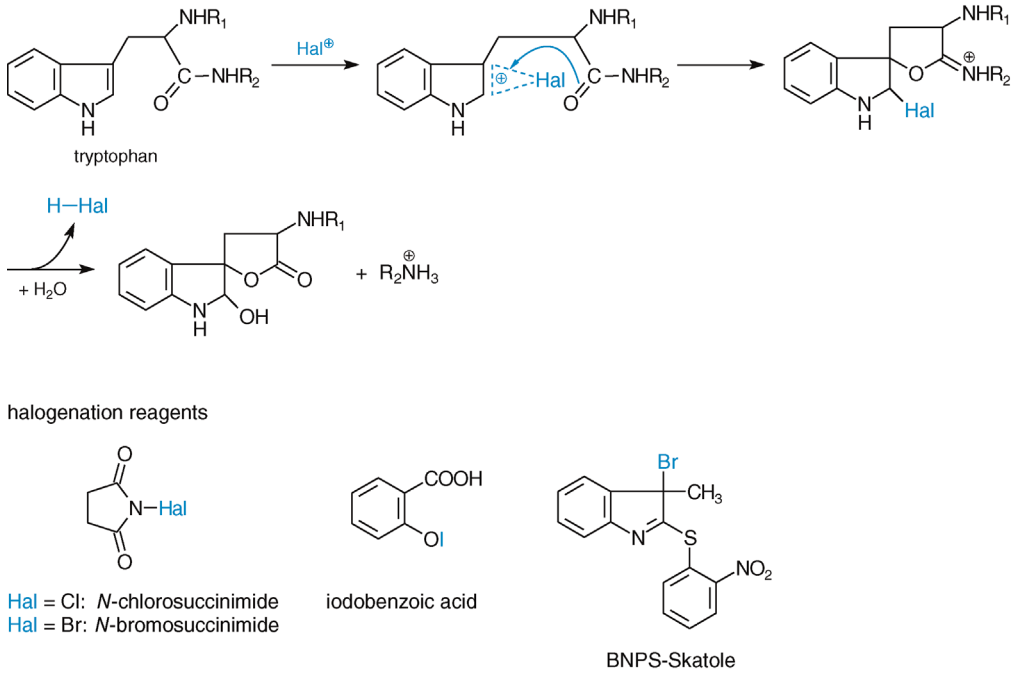


Figure 9.7 Cleavage at tryptophan via oxidative halogenation of the indole forming an oxindole.

Another partial hydrolysis occurs under extreme conditions at Xaa-Ser and Xaa-Thr (11 M HCl, 4 days), a method that in practice has little relevance.

Cleavage at Tryptophan As tryptophan is a rare amino acid it is also suitable for the generation of large fragments of a protein. Several reagents have been described by which to cleave a peptide bond where tryptophan is involved. Only *N*-bromosuccinimide (NBS), *N*-chlorosuccinimide (NCS), 3-bromo-3-methyl-2-(2-nitrophenylthio)-3*H*-indole (BNPS-skatole), and iodobenzoic acid are selective enough to allow for cleavage yields of up to 80%. These reagents have in common a positive polarized, and therefore electrophilic, halogen. Side reactions and additional cleavage sites at histidine and tyrosine residues often cannot be avoided. The cleavage at tryptophan is based on oxidative halogenation of the indole forming an oxindole ring (Figure 9.7).

Cleavage of Asn-Gly Peptide Bond using Hydroxylamine An Asn-Gly peptide bond within a protein sequence can spontaneously rearrange to an isopeptide bond. The starting point of this modification (Figure 9.8) is the formation of a succinimide ring of the carbonyl group of asparagine with the neighboring amino group of the glycine residue accompanied by a deamination of asparagine. Isomerization as well as racemization of asparagine is an often described reaction in protein chemistry. This succinimide ring is converted after its opening into aspartic acid or isoaspartic acid.

Isopeptide bonds during Edman sequencing result in termination of the reaction. But the formation of a succinimide can be used for a cleavage reaction with hydroxylamine, where, via a nucleophile attack, the protein is cleaved to form aspartic acid β-hydroxamate and a new N-terminal glycine residue.

Protein Sequence Analysis, Chapter 14

9.7 Summary

The field of application of protein fragmentation has changed in recent years. It is no longer necessary to cleave a protein with numerous different enzymes or reagents to identify even the last amino acid with classical protein chemical methods. In most cases the information of a protein sequence is already available from the DNA sequence. Nowadays, in protein analysis and proteomics, proteolytic cleavages are employed to identify proteins using peptide mass finger print (PMF) and sequencing of proteins and peptides using MS-MS technologies.

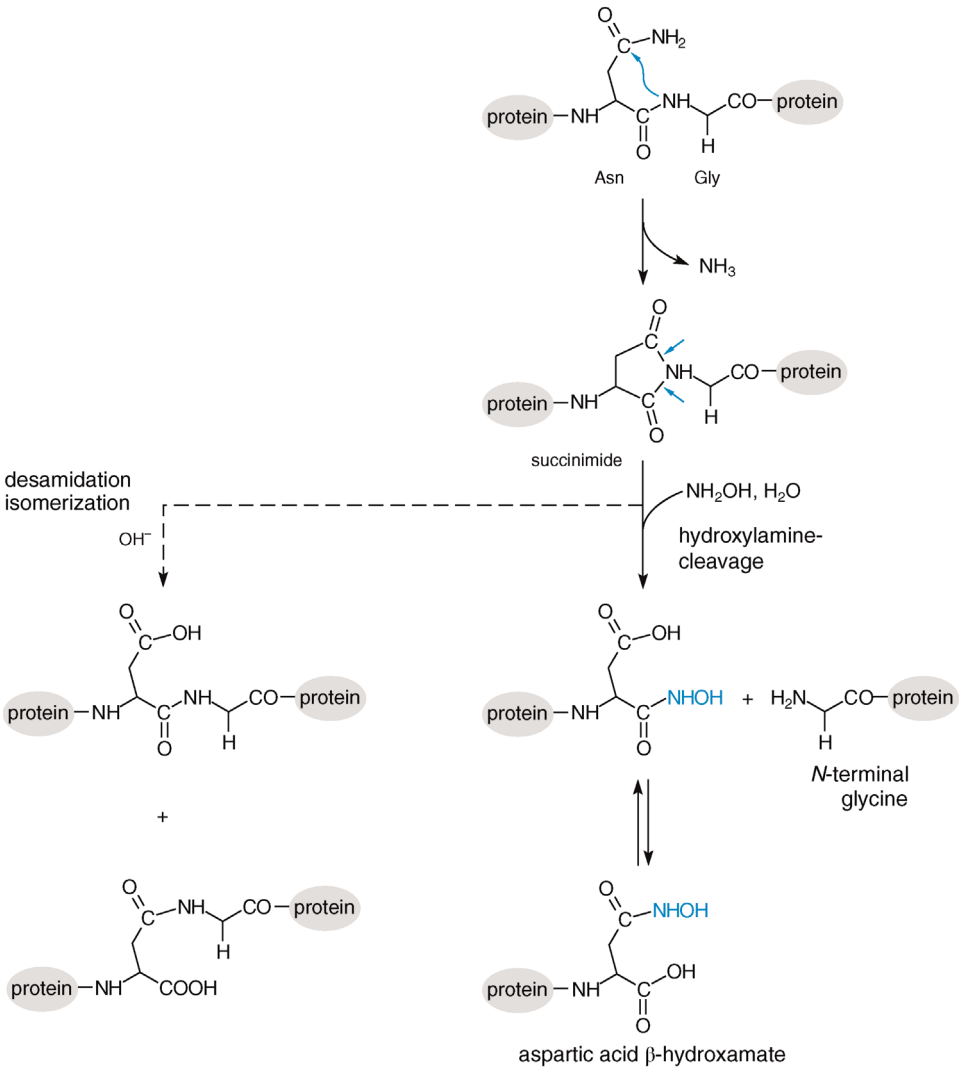


Figure 9.8 Reaction scheme of the hydroxylamine cleavage of Asn-Gly. The carboxyl group of Asn reacts with the amino group of the neighboring peptide bond to form a succinimide intermediate. The Asn-Gly bond is then cleaved by a nucleophilic attack of hydroxylamine to give an aspartic acid β-hydroxamate and a new N-terminal glycine.

Therefore, in most cases only a few highly specific and reproducible cleaving enzymes, like trypsin, endoproteinase Lys-C, and endoproteinase Glu-C, are used routinely.

Further Reading

Barret, A.J., Rawlings, N.D., and Woessner, J.F. (eds) (2013) *Handbook of Proteolytic Enzymes*, 3rd edn, Academic Press.

Patterson, S.D. (1994) From electrophoretically separated proteins to identification strategies for sequence and mass analysis. *Anal. Biochem.*, **221**, 1–15.

Sterchi, E. and Stöcker, W. (eds) (1999) *Proteolytic Enzymes; Tools and Targets*, Springer, Heidelberg.

Smith, B.J. (2003) Enzymatic cleavage of proteins, in *Protein Sequencing Protocols* (ed. B.J. Smith), Methods in Molecular Biology, vol. 211, pp. 49–61, Humana Press, Totowa, NJ.

Smith, B.J. (2003) Chemical cleavage of polypeptides, in *Protein Sequencing Protocols* (ed. B.J. Smith), Methods in Molecular Biology, vol. 211, pp. 63–82, Humana Press, Totowa, NJ.

Walker, J.M. (ed.) (2009) *The Protein Protocols Handbook*, Humana Press, Totowa, NJ.

Chromatographic Separation Methods

10

Reinhard Boysen

School of Chemistry, Monash University, Monash, Victoria 3800, Melbourne, Australia

Over the past three decades, high-performance liquid chromatography (HPLC) has become an indispensable tool for the separation, purification, and characterization of chemical and biological molecules. Despite this progress, the use of HPLC for the isolation of naturally occurring peptides or proteins from complex biological matrices and for their subsequent unambiguous structural elucidation still presents considerable challenges, for example, in the fields of peptidomics, proteomics, and degradomics.

Since proteins can occur as structural isoforms (as the result of chemical or biological post-translational modifications) or protein variants (from genetic modifications), their identification requires mass spectrometry (MS) and bioinformatics analysis. Such techniques may also be required for the purification of synthetic peptides and recombinantly produced proteins. Further structural elucidation can be performed with multiple-stage mass spectrometry, that is, with an ion-trap mass spectrometer, quadruple, high-resolution/orbitrap mass spectrometer, or Fourier transform mass spectrometer, or with proton, carbon, or heteronuclear NMR-spectroscopy.

Mass Spectrometry, Chapter 15

To develop a purification strategy tailored to a target peptide or protein, an in-depth knowledge of the available instrumentation, of the fundamental terms and concepts in chromatography, and of the basic physicochemical characteristics of peptides and proteins is required. These aspects are considered in Sections 10.1–10.3. The major chromatographic modes for the isolation and purification of peptides and proteins are described in Section 10.4. To take full advantage of a specific HPLC mode and to utilize time and resources effectively, a comprehensive method development is always recommended. Guidelines for method development in HP-RPC, the major chromatographic mode employed in peptide and protein analysis, are given in Section 10.5. Since peptides and proteins are generally isolated from complex biological matrices, often more than one chromatographic step is needed for their purification. The successive application of several suitable different chromatographic modes must consider their applicability to the protein of interest and their compatibility with each other and with the detection procedures. The concepts and implementations of two- or higher-dimensional separation schemes for peptide and protein purification are discussed in Section 10.6.

10.1 Instrumentation

The configuration of the chromatographic system optimally suited for a particular application is of considerable importance in the field of bioanalysis. Modern instrumentation provides many possibilities to configure a separation system using individual modules or alternatively a compact, preconfigured chromatographic system. A simple LC system consists of a binary pump, an injection system, a column, and a detector. The mobile phase can be pumped through the column at atmospheric pressure (open systems) or at elevated pressure, in medium, high, or

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

ultrahigh pressure systems. The development of high-performance liquid chromatography (HPLC) resulted in significant advances in speed, resolution, and automation and it has advanced to an important chromatographic method in analytical biochemistry. According to need, additional components can be added to a chromatographic system, including an eluent degasser, a (thermostated) autosampler, a column oven with integrated valves for column switching, a fraction collector, and so on. The individual components can be controlled manually or with a personal computer, which can be used to record and analyze the chromatograms. A chromatographic system can be tailored for a particular application, for example, for high-throughput (which may require a robot that loads the autosampler), for method development, for trace analysis (which may require a fluorescence detector), for two-dimensional separation (which requires an additional pump and switching valve), or for preparative purification, whereby the individual system components are integrated accordingly.

Three recent developments have had an immense impact on the field of bioanalysis: The coupling (hyphenation) of liquid chromatography (LC) with mass spectrometry, the miniaturization of separation columns, and the general availability of bioinformatics. The LC/MS-coupling enables the high-sensitive detection of molecular masses of peptides and proteins, the application of tandem MS procedures for the identification of peptides and associated post-translational modifications, and the *de novo* sequencing of peptides as well as the relative (or absolute) quantification of peptides directly after a chromatographic separation. After the chromatographic separation, the eluate from the column is infused into an electrospray ionization (ESI) mass spectrometer were the separated peptides and proteins are sequentially desolvated, ionized, led through the ion optics, and then detected with a mass analyzer. The optimum utilization of the electrospray technique requires a constant and low flow rate of the infused eluent stream, which can be conveniently produced by *micro-* or *nano LC* systems. Since the mass spectrometer is a concentration dependent detector, a decrease in the LC flow rate results in an increase in detection sensitivity. Furthermore, the utilization of a micro-spotter allows the deposition of eluate (and matrix) on sample plates suitable for subsequent matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI TOF MS). The development of analytic *micro-* and *nano-LC,* which can be categorized as capillary-LC, and the decreasing flow rates and column diameter reflect a trend of increasing *miniaturization,* which has also resulted in the development of HPLC chips with integrated enrichment and separation column. The individual components of these capillary HPLC systems are designed for specific tasks. The micro- or nano-LC pumps are designed for very small flow rates, and the diameter of the connecting capillary tubing, valve volumes, and detector cell volumes are adapted accordingly. For complex samples, as they occur in proteomics, the HPLC systems are necessarily multidimensional, which means they integrate two or more different chromatographic separation modes, for example, ion exchange chromatography and reversed-phase chromatography. The analysis of mass spectroscopic data from such a multidimensional HPLC system requires access to modern bioinformatics software and sequence data banks.

Mass Spectrometry, Chapter 15

10.2 Fundamental Terms and Concepts in Chromatography

Chromatography can be defined as a separation method, whereby the soluble components of a sample are distributed between two phases, one of which is *stationary* whilst the other, the mobile phase, moves either in an axial or radial direction depending on the column design features, with the analytes migrating in a predetermined, definite direction.

A chromatographic separation begins with the injection of the sample. The individual analytes, carried by the mobile phase, migrate through the chromatographic column with differing velocity, depending on the elution methods employed, and are detected, for example, by UV spectroscopy or mass spectrometry, in the order in which they leave the column. The resulting chromatogram plots the response of a detector to the concentration of the compound (peak) versus the length of time it takes to elute it from the column (retention time) (Figure 10.1). The time an unretained compound needs to migrate through the column is the column void time t_0. It is related to the void volume V_0, which is the sum of the interstitial volume between the particles of the stationary phase and the available volume within the particle pores. The elution

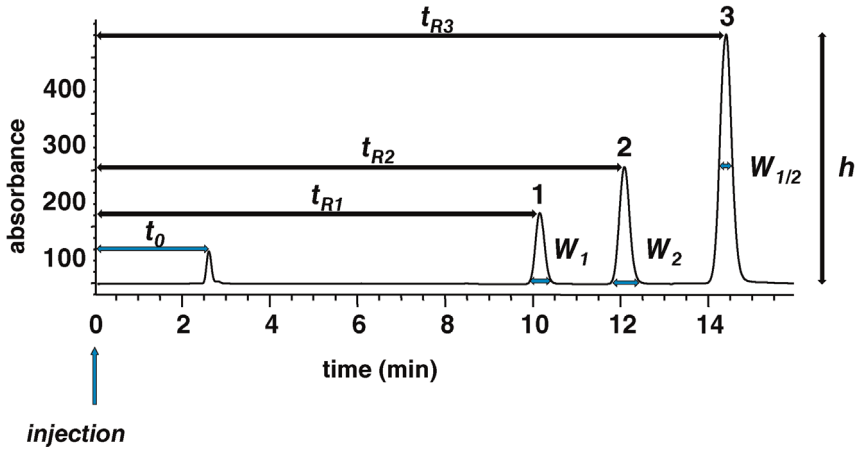


Figure 10.1 Schematic depiction of a chromatogram. Abbreviations: h = peak height, t_0 = column void time (time from injection to detection of unretained compound), t_{R1}, t_{R2}, t_{R3} = retention times (time from injection to detection of retained compounds), w = peak width (measured at baseline) and $w_{½}$ = peak width at half height (measured to characterize column efficiency).

of sample components is delayed through their interaction with the stationary phase, which is described by the retention time t_R or by the retention volume V_R.

The retention factor k is used to describe the retention independent of the column dimensions or flow rate:

$$k = \frac{t_R - t_0}{t_0} \tag{10.1}$$

where t_R is retention time and t_0 is column void time. Alternatively, the retention factor can be expressed in terms of elution volumes, since retention times, t_R and t_0, are related to the elution volumes and the flow rate F of the chromatographic system through the relationships:

$$V_R = t_R F \text{ and } V_0 = t_0 F \tag{10.2}$$

hence:

$$k = \frac{V_R - V_0}{V_0} \tag{10.3}$$

Thus, the retention factor relates to the number of additional column volumes beyond V_0 required to elute a compound. The retention factor can have values between $k = 0$ (no retention) and $k = \infty$ (irreversible adsorption) with values of 1–20 most preferred for practical and economic reasons. The retention factor can also be defined as the ratio n_s/n_m, where n_s is the total number of moles of the solute associated with the stationary phase and n_m is the total number of moles of the solute in the mobile phase:

$$k = \frac{n_s}{n_m} \tag{10.4}$$

To resolve two components, their retention factors must be different. The selectivity α of a chromatographic system describes the ability of a separation column to separate two compounds (1 and 2) based on their different retention factors k_1 and k_2. For a selectivity of $\alpha = 1$ no separation is possible. The selectivity is defined as:

$$\alpha = \frac{k_2}{k_1} \tag{10.5}$$

To evaluate the quality of a separation not only the peak distance between the two components must be considered but also their respective peak width. The resolution, R_S, of two adjacent peaks in a chromatogram is defined by the ratio of peak distance and their peak widths:

$$R_s = \frac{t_{R_2} - t_{R_1}}{\frac{1}{2}(w_1 + w_2)} \tag{10.6}$$

with the retention times t_{R1} and t_{R2} of two adjacent peaks and the respective peak widths w_1 and w_2. Two peaks can be un-separated ($R_S < 1$), partially overlapping ($R_S = 1$), or baseline separated ($R_S > 1.5$).

Dispersion effects of solutes in the chromatographic system are one cause of band broadening. The extent of band broadening is reflected in the column efficiency, which is usually expressed as the plate number, N or as the plate height, H (also called height equivalent to one theoretical plate, HETP):

$$N = \frac{L}{H} \tag{10.7}$$

where L is the column length.

The concept of "theoretical plates" goes back to the number of distillation plates present during fractionating distillation. The higher the number of theoretical plates at a particular column length L, the better the quality of column and the narrower the peaks.

The value of the plate number N is dependent on various chromatographic and solute parameters including the column length, L, the chromatographic particle diameter, d_p, the linear flow velocity, u, and the solutes' diffusivities (D_m and D_s) in the bulk mobile phase and within the stationary phase, respectively. The plate number can be defined as:

$$N = \left(\frac{t_R^2}{\sigma_t^2}\right) \text{ or } N = 16\left(\frac{t_R}{w}\right)^2 \tag{10.8}$$

where t_R is the retention time, and the peak variance of the eluted zone in time units. For practical convenience, it is often replaced with the peak width w. For Gaussian peaks w approximately corresponds to 4σ (4× peak standard variation).

To permit a comparison of column efficiencies of columns with identical bed dimensions packed with sorbent particles of different physical or chemical characteristics (e.g., different average diameter, ligand type), the plate height H is defined through the reduced plate height, h, while the linear flow velocity $u = L/t_0$ can be defined in terms of reduced mobile phase velocity, ν:

$$h = \frac{H}{d_P} \text{ and } \nu = \frac{u\, d_P}{D_m} \tag{10.9}$$

where L is the column length, d_P is the particle diameter, and D_m the diffusivity of the solute in the mobile phase.

The peak zone band broadening that is caused by various mass transport effects in the column is described through the dependency of the reduced plate height h on the velocity, u, or the reduced velocity ν, through the van Deemter–Knox equation:

$$h = A + \frac{B}{u} + Cu \text{ or } h = A\,\nu^{\frac{1}{3}} + \frac{B}{\nu} + C\nu \tag{10.10}$$

Here the A term expresses the Eddy diffusion and mobile phase mass transfer effects and is a measure of the packing quality of the chromatographic bed (and is constant for a given column), the B term entails the longitudinal molecular diffusion effects, while the C term incorporates mass transfer resistances within the stationary phase microenvironment, which describes the interaction of the solutes with the stationary phase.

To achieve optimal separation performance, the reduced plate height, h, needs to be as small as possible. The optimal flow velocity can be deduced from the minimum in h in the van Deemter–Knox plots (Figure 10.2).

The Deemter-Knox equation can be used to mathematically describe the chromatographic behavior of small molecular weight compounds like peptides to a good approximation, whereas for large molecules like proteins it has limitations, since the proteins vary considerably in shape and surface properties. As a consequence, the chromatographic effects can only be described with approximations and average values.

Figure 10.2 documents the difference between small and large molecules using the van-Deemter-plots of cytosine (M = 111) and myoglobin (M = 17 000) as an example. The plate height at different flow rates was determined with size exclusion chromatography. The plots reveal that:

1. the behavior of small molecules is determined by their diffusion; however, for large molecules the influence of the B term (longitudinal molecular diffusion) is negligible, particularly at higher flow velocities;

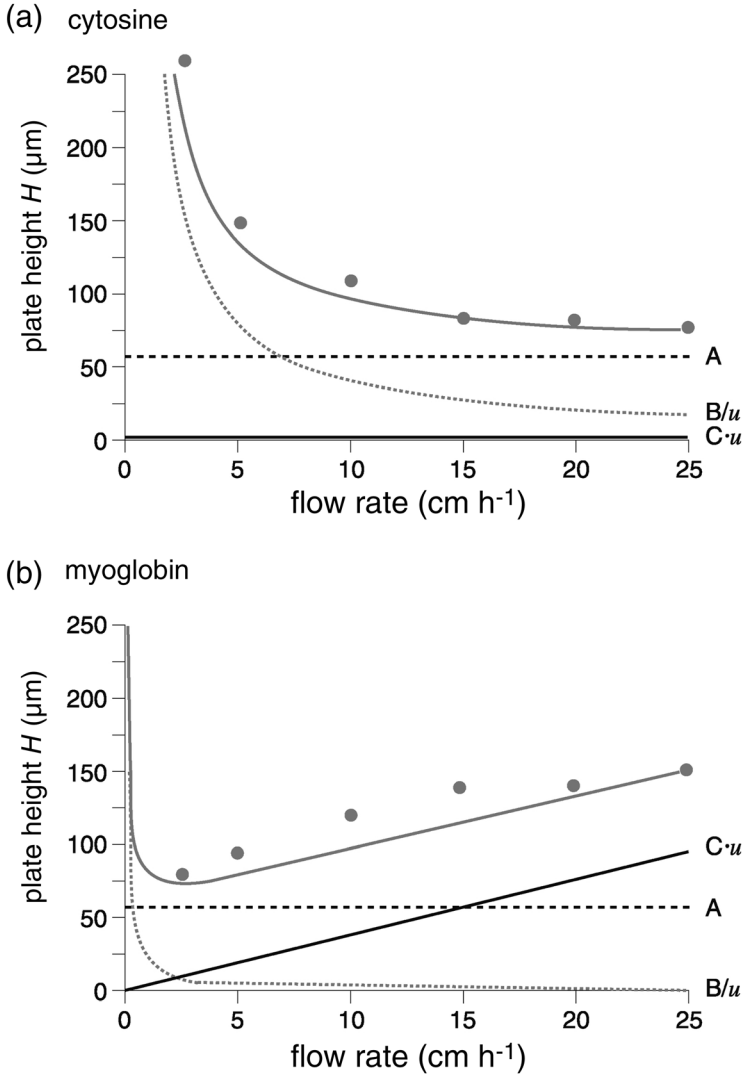


Figure 10.2 Van-Deemter-Knox plots for molecules of different molecular size (solid circles) and the individual contributions of the *A* term, *B/u* term and the *Cu* term to the plate height: (a) cytosine ($M = 111$) and (b) myoglobin ($M = 17\,000$). Source: adapted from Janson, J.C. and Ryden, L. (1989) *Protein Purification*, VCH, Weinheim.

2. the interaction with the stationary phase (C-term) of large molecules results in band broadening;
3. the optimal plate height for large molecules can be obtained at lower flow velocities.

Besides zone band broadening within the column, the band broadening can also arise from extra-column band broadening due to instrument characteristics:

$$\sigma_t^2 = \sigma_{column}^2 + \sigma_{extra}^2 \tag{10.11}$$

where σ_{column}^2 and σ_{extra}^2 are the peak variances arising due to column and extra-column effects, respectively. Careful attention must be paid to possible sources of extra-column effects (e.g., the choice of tubing in terms of length and diameter, the type of fittings, frits, the choice of detector cell volume) that can reduce the impact of σ_{extra}^2 on the overall h value.

The zone band broadening or peak dispersion, expressed as reduced plate height h, arises from kinetic, time-dependent phenomena. In the absence of secondary effects (e.g., slow chemical equilibria, pH effects, conformational changes) that could influence the chromatographic process, the resolution R_S can be expressed as:

$$R_S = \frac{1}{4} N^{\frac{1}{2}} (\alpha - 1) \left(\frac{k}{1 + k} \right) \tag{10.12}$$

This equation links three essential parameters that determine the quality of a chromatographic separation, namely, the retention factor k, the selectivity α, and the plate number N and,

therefore, describes the extent to which the zone spreading may cause the loss of separation performance. As will be demonstrated further below, this equation can be used to guide systematic method development for resolution optimization.

10.3 Biophysical Properties of Peptides and Proteins

The primary structure (amino acid sequence) and the folded structure (i.e., the secondary, tertiary, and quaternary structure) are the essential features of a peptide or protein, around which a chromatographic separation is designed. The 20 naturally occurring l-α-amino acids found in peptides and proteins vary with respect to the properties of their side chains. This chemical diversity is further increased where some of these side chains have been post-translationally modified with various chemical and biological modifications (e.g., acetylation, deamidation, glycosylation, lipidation, and phosphorylation). The side-chains are generally classified according to their polarity (e.g., nonpolar or hydrophobic versus polar or hydrophilic). The polar side chains are divided into three groups: uncharged, positively charged or basic, and negatively charged or acidic side chains. All N- and C-terminally unblocked peptides and proteins contain several ionizable basic and acidic functionalities. They, therefore, typically exhibit characteristic isoelectric points with the overall net charges and polarities in aqueous solutions varying with pH, solvent composition, and temperature. Cyclic peptides without ionizable side chains have a zero net charge, and they represent an exception. The number and distribution of charged groups influences the polarizability and ionization status of a peptide or protein, as well as the hydrophobicity. Table 10.1 lists the characteristic data for the most common l-α-amino acids found in peptides and proteins and gives a summary of the N- and C-terminal groups. These data can be used to guide the selection of the optimal separation conditions for the resolution of peptide and protein mixtures. They can be used to evaluate the impact of amino acid composition (which, for example, directs the choice of eluent composition or the gradient range in RP-HPLC) on retention behavior. They can also point to the impact of amino acid substitution or deletion with small peptides on retention; or alternatively can be used to guide the identification of peptide fragments derived from proteolytic (chemical or enzymatic) digestion of proteins.

NMR, Chapter 18

In solution, a polypeptide or protein can, in principle, explore a relatively large array of conformational space. For small peptides (e.g., less than 15 amino acid residues) a defined secondary structure (α-helical, β-sheet, or β-turn motif) is generally absent. With increasing polypeptide chain length, depending on the nature of the amino acid sequence, specific regions/domains of a polypeptide or protein can adopt preferred secondary, tertiary, or quaternary structures. In aqueous solutions this folding, which internalizes the hydrophobic residues and thus stabilizes the polypeptide structure, becomes a significant feature of peptides and proteins for chromatographic separations. A critical factor to be considered in the selection of a HPLC procedure is that the experimental conditions will inevitably cause perturbations of the conformational status of these biomacromolecules. In most cases, an integrated biophysical experimental strategy (including ^{1}H 2D NMR, FTIR, CD-ORD spectroscopy, and ESI-MS) is required to determine the secondary and higher order structure of a polypeptide or protein in solution or in the presence of specific ligands or of possible self-self-aggregation effects with peptides or proteins that may have occurred during the HPLC separation.

IR Spectrometry, Section 7.4

ORD and CD, Section 7.7.2

Mass Spectrometry, Chapter 15

The peptide bond absorbs strongly in the far ultraviolet (UV) region of the spectrum ($\lambda = 205$–215 nm). Hence UV detection is the most widely used method for detection of peptides and proteins in HPLC. Besides absorbing in the far UV range, the aromatic amino acid residues also absorb light above 250 nm, owing to the delocalized π-systems of the aromatic residues. Knowledge of the UV-spectra, in particular the extinction coefficients of the non-overlapping absorption maxima of these amino acids, allows, in conjunction with UV-diode array detection (UV-DAD) and second derivative or difference UV-spectroscopy, verification of peak purity and determination of the aromatic amino acid content of peptides and proteins. Knowledge of the relative UV/VIS absorbancy of a peptide or protein is thus crucial, since the choice of detection wavelength of peptides and proteins in RP-HPLC (and in the other HPLC modes) depends on the different UV cut-offs of the eluents used. The common use of $\lambda = 215$ nm as the preferred detection wavelength for most analytical applications with peptides

UV/VIS Spectroscopy, Section 7.2

Table 10.1 Properties of the common L-α-amino acid residues and termini.

Three-letter code	One-letter code	Mono-isotopic mass[a] (amu)	Partial specific volume[b] (Å^3)	Accessible surface area[c] (Å^2)	pK_a of side-chain[d] or termini[e]	Relative hydrophobicity[f]
Ala	A	71.03711	88.6	115		0.06
Arg	R	156.10111	173.4	225	12.48	–0.85
Asn	N	114.04293	117.7	160		0.25
Asp	D	115.02694	111.1	150	3.9	–0.20
Cys	C	103.00919	108.5	135	8.37	0.49
Gln	Q	128.05858	143.9	180		0.31
Glu	E	129.04259	138.4	190	4.07	–0.10
Gly	G	57.02146	60.1	75		0.21
His	H	137.05891	153.2	195	6.04	–2.24
Ile	I	113.08406	166.7	175		3.48
Leu	L	113.08406	166.7	170		3.50
Lys	K	128.09496	168.6	200	10.54	–1.62
Met	M	131.04049	162.9	185		0.21
Phe	F	147.06841	189.9	210		4.8
Pro	P	97.05276	122.7	145		0.71
Ser	S	87.03203	89	115		–0.62
Thr	T	101.04768	116.1	140		0.65
Trp	W	186.07931	227.8	255		2.29
Tyr	Y	163.06333	193.6	230	10.46	1.89
Val	V	99.06841	140	155		1.59
α-amino					7.7–9.2	
α-carboxyl					2.75–3.	

a) Monoisotopic masses (amu) of N- and C-terminal groups: hydrogen (H) 1.00782, N-formyl (HCO) 29.00274, N-acetyl (CH_3CO) 43.01839. Free acid (OH) 17.00274, amide (NH_2) 16.01872.
b) Zamyatnin, A.A. (1972) *Progr. Biophys. Mol. Biol.*, **24**, 107–123.
c) Chothia, C. (1975) *Nature*, **254**, 304–308.
d) Dawson, R.M.C., Elliot, D.C., Elliot, W.H., and Jones, K.M. (1986) *Data for Biomedical Research,* 3rd edn, Clarendon Press: Oxford.
e) Rickard, E.C., Strohl, M.M., and Nielsen, R.G. (1991) *Anal. Biochem.,* **197**, 197–207.
f) Wilce, M.C.J., Aguilar, M-I., and Hearn, M.T.W. (1995) *Anal. Chem.,* **67**, 1210–1219.

and proteins is a good compromise between detection sensitivity and potential detection interference due to light absorption by the eluent. However, wavelengths between 230 and 280 nm are frequently employed in preparative applications, where the use of more sensitive detection wavelengths could result in overloading of the detector response (usually above an absorbance value of 2.0–2.5 AU). Three aromatic amino acids – phenylalanine, tryptophan, and tyrosine – also show fluorescence. Since the fluorescence of the folded/unfolded state of proteins can vary considerably, it can be used to monitor the change of their folding status. However, it needs to be considered that the intensity of the fluorescence is dependent on the solvational environment, whereby the intensity of the fluorescence decreases inversely proportionally with the polarity of the solvent.

10.4 Chromatographic Separation Modes for Peptides and Proteins

There are several modes of HPLC currently in use for peptide and protein analysis, namely, size exclusion chromatography, reversed-phase chromatography, normal phase chromatography,

Table 10.2 Chromatographic separation modes for peptides and proteins.

Chromatographic mode	Acronym	Exploited molecular properties
Size exclusion or gel permeation chromatography	SEC or GPC	Molecular mass, hydrodynamic volume
Reversed-phase chromatography	RPC	Hydrophobicity
Normal phase chromatography	NPC	Polarity
Hydrophilic interaction chromatography	HILIC	Hydrophilicity
Aqueous normal phase chromatography	ANPC	Hydrophilicity
Hydrophobic interaction chromatography	HIC	Hydrophobicity
Anion exchange chromatography	AEX	Net negative charge
Cation exchange chromatography	CEX	Net positive charge
Affinity chromatography	AC	Specific interaction
Immobilized metal ion affinity chromatography	IMAC	Complexation

hydrophilic interaction chromatography, aqueous normal phase chromatography, hydrophobic interaction chromatography, ion exchange chromatography, and affinity chromatography, with the latter including immobilized metal ion affinity chromatography and biospecific/biomimetic affinity chromatography. Their principles are explained below. These and also several less frequently used chromatographic modes, for example, hydroxyapatite chromatography, mixed mode chromatography, charge transfer chromatography, or ligand-exchange chromatography can be operated under isocratic (i.e., fixed eluent composition), step-gradient, or gradient elution conditions, which change eluent conditions either in variable steps or continuously with the exception of HP-SEC, which is usually only performed under isocratic conditions. All modes can be used in analytical, semi-preparative, or preparative situations. To achieve optimal selectivity and hence resolution of peptides and proteins in high-performance chromatographic separations, irrespective of whether the task at hand is of analytical or preparative nature, the choice of the chromatographic mode must be guided by the properties of the analytes (i.e., their molecular size/shape hydrophobicity/hydrophilicity, net charge, isoelectric point, solubility, function, antigenicity, carbohydrate content, content of free SH groups, exposed histidine residues, exposed metal ions). A list of chromatographic modes and the molecular properties of the target compounds that form the basis of each separation mode is given in Table 10.2.

In addition to the above-mentioned functional characteristics of these chromatographic systems, other chemical and physical parameters of the mobile and stationary phase impact on the resolution, mass recovery, and bioactivity preservation in separations of polypeptides or proteins during liquid chromatographic separations. These parameters are listed in Table 10.3.

Table 10.3 Chemical and physical factors of the mobile and stationary phase that contribute to variation in the resolution, mass recovery, and bioactivity preservation of polypeptides, proteins, and other biomacromolecules in HPLC.

Mobile phase	Stationary phase
Buffer composition	Particle size
Ionic strength	Particle size distribution
pH	Particle compressibility
Organic solvents	Surface area
Metal ions	Pore diameter
Chaotropic reagents	Pore diameter distribution
Oxidizing or reducing reagents	Ligand composition
Loading concentration and volume	Ligand density
Temperature	Surface heterogeneity

10.4.1 High-Performance Size Exclusion Chromatography

High-performance size exclusion chromatography (HP-SEC), also called high-performance gel permeation chromatography (HP-GPC), is performed with porous stationary phases and separates analytes according to their molecular mass or, more precisely, their hydrodynamic volume. The separation of analytes is based on the concept that molecules of different hydrodynamic volume (Stokes radius) permeate to different extents into porous HP-SEC separation media and thus exhibit different permeation coefficients according to differences in their molecular masses/hydrodynamic volumes. Analytes with a molecular weight larger than the exclusion limit (usually listed in the technical information provided by the column manufacturer) are excluded from the pores and elute in the void volume of the column. As a non-retentive separation mode HP-SEC is usually operated with isocratic elution using aqueous low salt mobile phases.

HP-SEC can be used for group separation or high-resolution fractionation. In the group separation mode, HP-SEC removes small molecules from large molecules and is also suitable for buffer exchange or desalting. In the high-resolution fractionation mode, HP-SEC separates various components in a sample due to their different hydrodynamic volumes or can be employed to perform a molecular weight distribution analysis.

As HP-SEC columns (ideally) have no adsorption capacity and dilute the sample upon elution, they are not normally used in the initial capture or for intermediate purification in multistep chromatographic processes, but are, however, suitable for final polishing, for example, for the final removal of unwanted aggregates or multimeric forms of the proteins or other impurities of significantly different molecular weight. HP-SEC can be performed directly after HP-AC, HP-HIC, or HP-IEX without a buffer exchange.

10.4.2 High-Performance Reversed-Phase Chromatography (HP-RPC)

High-performance reversed-phase chromatography (HP-RPC) is the most frequently used analytical mode for peptide and protein analysis. In reversed-phase chromatography, the polarity of the stationary and mobile phase is the reverse of that used in normal phase chromatography. Peptides and proteins are loaded onto the column under aqueous conditions and eluted with a mobile phase containing an organic solvent. The column contains a porous or nonporous stationary phase with immobilized nonpolar ligands.

HP-RPC separates compounds according to their relative hydrophobicity. The most commonly accepted theory of the retention mechanism in HP-RPC is based on the solvophobic theory, which describes the hydrophobic interaction between the nonpolar surface regions of the analytes and the nonpolar ligands of the stationary phase. According to the solvophobic theory, the interaction of peptides and proteins in reversed-phase chromatography occurs through solvophobic exclusion of the solute from the aqueous mobile phase and the binding to the nonpolar surface of the immobilized ligand (solvophobic effect). The retention then depends on the size of the contact area ΔA of the hydrophobic surface of the solute and the hydrophobic area of the immobilized ligand as well as on the surface tension (γ) of the eluents. The surface tension γ depends on the composition of the mobile phase (volume fraction of the organic solvent, for example, acetonitrile, expressed as % B), the parameter that is varied, or more precisely lowered, during gradient elution:

$$\ln k = A + \frac{N_A}{RT\Delta A\gamma} \tag{10.13}$$

where: k is retention factor,
A is a constant,
N_A is the Avogadro constant,
R is the gas constant,
T is the absolute temperature,
and ΔA and γ are defined above.

Typically, the nonpolar ligands are immobilized onto the surface of spherical, porous, or nonporous silica particles, although nonpolar polymeric sorbents (e.g., those derived from

crosslinked polystyrene–divinylbenzene) can also be employed. Silica-based packing materials of 3–10 µm average particle diameter and 70–1000 Å pore size, with *n*-butyl, *n*-octyl, or *n*-octadecyl ligands, are widely used for the separation of peptides and proteins. Silica particles of 1 µm to more than 65 µm in diameter have been developed in various size distributions and configurations (e.g., spherical, irregular, with various pore geometries and pore connectivities, and in pellicular, fully porous or monolithic structures) by various routes of manufacture and with different silica types. They are grouped into type I, type II, or type III silica according to purity and metal content. For low molecular mass (<4000 Da) polypeptides, silica materials of 70–80 Å pore size and 3–5 µm average particle diameter are often used, which maximizes loading capacity and retention. For proteins in the mass range 4000–500 000 Da a pore size of 300 Å allows maintenance of high efficiency, whereby the loading capacity can be increased by increasing the column diameter. Macro-porous HP-RPC columns of 1000 Å pore size are increasingly used for the fractionation of very complex proteins samples.

In HP-RPC, an organic solvent (i.e., methanol, ethanol, acetonitrile, *n*-propanol, tetrahydrofuran) is used as a surface tension modifier in the chromatographic eluent, which has a particular elution strength, viscosity, and UV cut-off. Mobile phase additives, that is, acetic acid (AA), formic acid (FA), trifluoroacetic acid (TFA) and heptafluorobutyric acid (HBFA), are used to obtain a particular pH value, typically at low pH (e.g., pH ≈ 2 for silica based materials), with the exception of polymeric stationary phases, which have an extended pH range of pH 1–12. Some mobile phase additives may also function as ion pair reagents, which interact with the ionized analytes to form overall neutral eluting species and also suppress silanophilic interactions between free silanol groups on the silica surface and basic functional groups of the analytes. The properties of the additives determine their suitability for use with electrospray ionization mass spectrometry. Strong ion pair interactions between analytes and mobile phase additives can suppress the ionization of the analytes in electrospray ionization mass spectrometry (ESI-MS).

Mass Spectrometry, Chapter 15

HP-RPC can be operated in the isocratic, step gradient, or the continuous gradient elution mode and is frequently used as an intermediate or final polishing step in a multistep purification. It is ideally positioned after HP-IEX because it allows desalting and the separation of the sample in a single step.

Since most peptides and proteins possess some degree of hydrophobicity, HP-RPC techniques dominate the separation of peptides and proteins at the analytical and semi-preparative levels.

10.4.3 High-Performance Normal-Phase Chromatography (HP-NPC)

Chromatographic systems in which the stationary phase is more polar than the mobile phase had been developed at the beginning of the modern era of liquid chromatography and were known under the acronym "normal phase" liquid chromatography. High-performance normal-phase chromatography (HP-NPC) can be performed on unmodified silica and separates analytes according to their intrinsic polarity. The retaining mobile phase contains less polar organic solvents and the eluting mobile phase consists of more polar organic solvents. Water, due to its extreme polarity, adsorbs to most NP stationary phases, and significantly affects the separation reproducibility. In contrast to HP-RPC with immobilized *n*-alkyl ligands, where the interaction of solute and stationary phase is based on solvophobic phenomena, the interaction in HP-NPC is based on adsorption. The retention behavior of peptides and proteins in HP-NPC is often described in terms of the classical concepts inherent to multisite displacement and site occupancy theory.

HP-NPC is mainly used for the separation of, for example, polyaromatic hydrocarbons (PAHs), heteroaromatic compounds, nucleotides, and nucleosides, and much less frequently for protected synthetic peptides, deprotected small peptides in the "flash" chromatographic mode, and protected amino acid derivatives used in peptide synthesis. Originally, HP-NPC was limited to unmodified silica columns; however, recent work has utilized polar bonded phases such as amino (-NH_2), cyano (-CN), or diol (-COHCOH-) coated sorbents. Such modified normal phase packing materials were suitable for polar bonded phase chromatography (PBPC), which was used for the separation of peptides and proteins. Today, one of the main applications of modified normal phase silica materials is their use in HPLC-integrated solid-phase extraction (SPE) procedures. These types of sorbents, particularly when used as pre-column packing materials in LC-LC

column switching settings, in conjunction with restricted access sorbents materials (RAMs), allow multiple injections of untreated complex biological samples, for example, hemolyzed blood, plasma serum, fermentation broth, and cell tissue homogenates, for the isolation of bioactive peptides. Typically, with RAMs, hydrophilic, electroneutral diol groups are immobilized onto the outer surface of spherical particles. This layer prevents non-specific interactions between the support matrix and protein(s) or other high molecular weight biomolecules, which are thus excluded from the interior regions of the particle and elute as non-retained components. The inner surfaces of the porous RAM particles are, however, chemically modified with *n*-alkyl ligands, which are only freely accessible for low molecular weight analytes, such as peptides. As a consequence, significant enrichment or partial resolution of peptide analytes can be achieved. HP-NPC can be operated in isocratic, step gradient, or gradient elution mode.

10.4.4 High-Performance Hydrophilic Interaction Chromatography (HP-HILIC)

High-performance hydrophilic interaction chromatography (HP-HILIC) is performed on porous stationary phases with immobilized hydrophilic ligands and separates analytes according to their hydrophilicity. This variant of the HP-NPC mode was introduced by Alpert in 1990, based on polyaspartic acid immobilized onto silica and used for the separation of amino acids, small peptides, and simple malto-glycosides with mobile phases of high organic solvent content. A pseudo-HILIC separation of simple saccharides was performed on a BondaPak-NH_2 column as early as 1975. The HP-HILIC mode has since been applied for the separation of various analytes, that is, simple carbohydrates and amino acids as well as peptides.

In HILIC, polar sorbents with amide-, aminopropyl-, cyanopropyl-, diol-, cyclodextrin-, poly (succinimide)-, and sulfoalkylbetaine phases are employed, and the non-aqueous mobile phases of NPC are replaced with high-organic, low-aqueous eluents. The initial mobile phase has a high organic solvent and a low water content. Elution of compounds from HP-HILIC columns is achieved by increasing the water content in the mobile phase. The elution order was initially thought to be more or less opposite to that seen in HP-RPC separation. Although it intuitively would seem that retention in HILIC would simply be the "reverse" of that in HP-RPC, studies on the orthogonality of separations in two-dimensional liquid chromatography have demonstrated that HP-HILIC (with a bare silica sorbent) and HP-RPC can be a suitable combination for proteomic analysis in 2D systems. Compared to the non-aqueous (organic) mobile phase in HP-NPC, the partly aqueous mobile phase used in HP-HILIC allows greater solubility of many polar and hydrophilic compounds and fast separation of polar compounds can be achieved due to the low viscosity of the highly organic mobile phase. Moreover, the high content of organic solvent in the mobile phase favors the ionization of polar compounds in subsequent ESI-MS, and thus provides enhanced detection sensitivity for these compounds.

Considerable scientific debate exists about the physical basis of the separation mechanism in HP-HILIC. The roles of thermodynamic or kinetic effects in controlling the resolution and separation efficiencies have yet to be fully explored. Although it was proposed that the retention of polar compounds in HILIC is through partitioning between the bulk of the mostly organic mobile phase and a stagnant water-enriched layer semi-immobilized on the surface of silica, processes of partitioning or adsorption or combinations of it have also been suggested to be responsible for generating retention in HP-HILIC.

Mixed mode HILIC/cation-exchange chromatography (HILIC/CEX) is another way to separate peptides. The contact region concept developed for HP-RPC can be used to rationalize the retention of amphipathic α-helical peptides, for example, based on experiments with a poly (2-sulfoethyl aspartamide)-silica (PolySulfoethyl A) strong cation-exchange column in the HILIC/cation-exchange chromatography (HILIC/CEC) mode and with a Zorbax SB300-C8 reversed-phase column for separating amphipathic α-helical peptides. A substitution in the hydrophilic face of the peptide resulted in a substantive effect on retention in HP-HILIC/CEC; this was not observed in HP-RPC, whereas a substitution in the hydrophobic face of the peptide resulted in a distinct effect on retention in HP-RPC that was not observed in HP-HILIC.

HP-HILIC can be operated in isocratic, step gradient, or gradient elution mode, where the retaining mobile phase is organic and the eluting mobile phase is aqueous. Being more suited to the isolation of polar substances, HP-HILIC, when linked to electrospray mass

spectrometry, has mainly found application for analysis of phosphopeptides and glycopeptides.

10.4.5 High-Performance Aqueous Normal Phase Chromatography (HP-ANPC)

Recently, a new chromatographic mode, namely, high-performance aqueous normal phase chromatography (HP-ANPC) has been developed on stationary phases based on silica hydride surfaces. One unique feature of silica hydride stationary phases is their ability to be employed over a broad range of mobile phase compositions from 100% aqueous to pure organic solvents. A unique advantage of HP-ANPC is that it allows the separation of compounds with a broad range of hydrophilicities in the same mixture using predominantly water-rich mobile phase compositions.

The retention principle in HP-ANPC is analogous to that found HP-NPC but the mobile phase contains water as part of the binary eluent. The difference between HP-ANPC and HP-HILIC is that in HP-HILIC retention is determined by an adsorbed water layer on the surface that is either not present or substantially smaller on silica hydride surfaces employed in HP-ANPC. The precise separation mechanisms that operate in HP-ANPC are still subject to ongoing research and the interchangeable use of the terms HP-HILIC and HP-ANPC has led to some confusion in the literature. However, due to its versatility, HP-ANPC has been applied for the separation of peptides in isocratic and gradient elution modes.

10.4.6 High-Performance Hydrophobic Interaction Chromatography (HP-HIC)

High-performance hydrophobic interaction chromatography (HP-HIC) is used for protein separations involving hydrophobic sorbents. In HP-HIC the binding of proteins to the stationary phase occurs at high salt concentrations of the mobile phase and elution is performed through a decrease in salt concentration. In HP-HIC the two most widely used base matrices are hydrophilic carbohydrates, for example, crosslinked agarose or synthetic copolymer materials. The type of immobilized ligand (alkyl or aryl ligand) influences the selectivity of the stationary phase. In HP-HIC, nonpolar ligands with lower hydrophobicity and lower ligand density (ca. $1/10^{th}$ of to that of HP-RPC sorbents) are employed. These differences between HP-RPC and HP-HIC have fundamental effects on the recovery of proteins in the bioactive state, as well as on the selectivity of the system. The protein binding capacities of HP-HIC sorbents increases with increasing *n*-alkyl chain length at a given ligand density as well as with increasing ligand density, but will reach a plateau for very high ligand density levels. HP-HIC sorbents should be selected on the basis of the critical hydrophobicity concept.

HP-HIC separates proteins according to their hydrophobicity differences. The retention mechanism is based on the reversible interaction between a protein and a hydrophobic surface of a chromatographic support, depending on changes of the microscopic surface tension associated with the composition of the mobile phase. Similar to HP-RPC, where the decrease in surface tension of the eluent is achieved through an increase in the organic solvent content of the mobile phase, in HP-HIC this is achieved with a decreasing salt concentration, that is, by increasing the water content of the eluent.

The selectivity of protein separations in HP-HIC can be influenced through stationary phase parameters (such as type and density of the immobilized hydrophobic ligand type of base matrix) and mobile phase parameters (such as type and concentration of salt, the addition of organic modifiers or surfactants, and the pH) as well as the temperature of the column.

In HIC the hydrophobic interactions between polypeptides or proteins and a sorbent are influenced by the use of salts in the mobile phase (various salts have differing molal surface tension increment values, see Table 10.4), by the use of different types or concentrations of salt. The surface tension of the mobile phase, γ, can be related to the molal surface tension increment, σ, and the molal concentration, m, of the salt by the equation:

$$\gamma = \gamma^{\circ} + \sigma m \tag{10.14}$$

where $\gamma^{\circ} = 72\,\mathrm{dyn\,cm^{-1}}$ for water.

Table 10.4 Salts used in the mobile phase of HP-HIC.

Salt type	Molal surface tension increment σ ($\times 10^3$ dyn g cm^{-1}·mol^{-1})
Calcium chloride	3.66
Magnesium chloride	3.16
Potassium citrate	3.12
Sodium sulfate	2.73
Potassium sulfate	2.58
Ammonium sulfate	2.16
Magnesium sulfate	2.10
Sodium dihydrogen phosphate	2.02
Potassium tartrate	1.96
Sodium chloride	1.64
Potassium perchlorate	1.40
Ammonium chloride	1.39
Sodium bromide	1.32
Sodium nitrate	1.06
Sodium perchlorate	0.55
Potassium thiocyanate	0.45

Table 10.5 shows the parameters of surface tension increment σ, initial mobile phase concentration m, and surface tension γ of common aqueous salt buffers.

The minimum surface tension reached in the HP-HIC of polypeptides or proteins with binary water–salt systems corresponds to the surface tension of pure water, that is, 72 dyn cm^{-1}.

The effect of different salts on the hydrophobic interaction follows the (lyotropic) Hofmeister series for the precipitation of proteins from aqueous solution (Table 10.6). The Hofmeister series ranks the effect of anions and cations in promoting protein precipitation. Ions with higher salting-out effect promote binding to hydrophobic interaction sorbents, whereas ions with higher salting in-effect promote elution from hydrophobic interaction sorbents.

The salts at the beginning of the series promote hydrophobic interactions and protein precipitation (salting-out effect). These are called anti-chaotropic (or kosmotropic) salts and are considered to be water structuring. The addition of kosmotropic salts to the equilibration buffer and sample buffer promote protein immobilized ligand interaction in HIC. The salts at the end of the series (salting-in or chaotropic ions) randomize the structure of liquid water and tend to decrease the strength of hydrophobic interactions. The chloride anion is considered to be approximately neutral with respect to the water structure.

In general, the effect of the salt cations on the preferential interaction parameters is not as pronounced as the salt anions, in particular when the cation is monovalent; however, divalent cations tend to bind to proteins. Typically, kosmotropic (anti-chaotropic) salts (i.e., ammonium

Table 10.5 Parameters related to the surface tension of frequently used aqueous salt buffers.

Salt buffer type	σ ($\times 10^3$ dyn g cm^{-1}·mol^{-1})	m ($\times 10^3$ mol g^{-1})	γ (dyn cm^{-1})
Ammonium sulfate	2.16	2	77.31
Sodium chloride	1.64	2	76.29
Magnesium sulfate	2.1	1.4	75.95
Sodium sulfate	2.73	1	75.74
Sodium perchlorate	0.55	2	74.11
Sodium phosphate	2.02	0.05	73.01

Table 10.6 The Hofmeister series.

	← Salting-out effect (precipitation)
Anions	PO_4^{3-}, SO_4^{2-}, CH_3COO^-, Cl^-, Br^-, NO^{3-}, ClO_4^-, I^-, SCN^-
Cations	NH_4^+, K^+, Na^+, Cs^+, Li^+, Mg^{2+}, Ca^{2+}, Ba^{2+}
	Salting-in effect →

sulfate, sodium sulfate, magnesium chloride) of high molal surface tension increment are to be preferred in HP-HIC applications for polypeptides and proteins.

Since protein–hydrophobic surface interactions are enhanced by high ion strength buffer solutions, HP-HIC is a suitable next step after ammonium sulfate precipitation or HP-IEX elution with high-salt buffer. In combination with non-denaturing mobile phases, proteins can potentially be eluted in their native conformation from HP-HIC sorbents.

10.4.7 High-Performance Ion Exchange Chromatography (HP-IEX)

High-performance ion exchange chromatography (HP-IEX) is performed on stationary phases with immobilized charged ligands and separation occurs according to electrostatic interactions between the charged surface of the analyte(s) and the complementarily charged surface of the sorbent. In high-performance anion exchange chromatography (HP-AEX), peptides and proteins are separated according to their net negative charge, whereby the retaining mobile phase is aqueous, of high pH, and low salt concentration and the eluting mobile phase is either aqueous, of high pH and high salt concentration, or aqueous and of low pH. In contrast, high-performance cation exchange chromatography (HP-CEX) separates analytes according to their net positive charge, whereby the retaining mobile phase is aqueous, of low pH, and low salt concentration and the eluting mobile phase is either aqueous, of low pH and high salt concentration, or aqueous and of high pH.

Both weak and strong cation exchangers (e.g., based on carboxymethyl or sulfonopropyl ligands), as well as weak and strong anion exchangers (e.g., dimethylamino or quaternary ammonium ligands), are commercially available and are very suitable for the HP-IEX of peptides and proteins. For peptide and protein separations, the use of a strong cation-exchange column has a considerable advantage over other ion exchange separation modes, since this column can retain its negative charge character over a large pH range, from acidic to neutral. In peptides and proteins, at neutral pH, the side chain carboxyl groups of the acidic amino acid residues (glutamic acid and aspartic acid) are completely ionized. Below pH 3 they are almost completely protonated. A change of pH therefore allows the retention of peptides and proteins to be varied according to the modified net charge of these biosolute(s).

The "net charge" concept has been widely used as a predictive basis by which to anticipate the retention behavior of proteins with both anion and cation exchange stationary phases. According to this model, a protein will be retained on a cation-exchange column if the eluent pH is lower than the p*I* value of the protein, since under these conditions the protein will carry positive net charges. Conversely, a protein will be retained on an anion exchange column when the eluent pH is above the p*I* of the protein. Finally, with a mobile phase of a pH that is equal to the p*I* of the protein, the surface of the protein can be considered as electrostatically neutral and the protein should not be retained on either cation or anion exchange columns. This classical model is now considered as simplistic. Recent investigations have revealed that the magnitude of electrostatic interactions between a protein and the stationary phase surface in HP-IEX is dependent on the charge density of the stationary phase, the mobile phase composition, and on the number and distribution of charged sites on the protein molecule since they define its surface topography and electrostatic contact area with the stationary phase. As a consequence, a variation of chromatographic parameters can alter the affinity of the protein for the stationary phase through changes in the overall electrostatic surface charge or through specific electrostatic interactions of the displacer co-ions and counter-ions with surface charge groups on the protein or with the immobilized charged ligand. In addition, possible changes of the three-

dimensional structure of the proteins may have significant effects on protein retention. Studies on the influence of the experimental parameters on the number of charged interactive sites of the proteins involved in its binding to stationary phases have resulted in the development of the concept of an electrostatic interactive area (or ionotope) through which the protein is thought to bind to the stationary phase. Peptides and proteins can be separated in HP-IEX by isocratic, step, or by gradient elution at high resolution and with high capacity.

10.4.8 High-Performance Affinity Chromatography (HP-AC)

High-performance affinity chromatography (HP-AC) is performed on stationary phases containing immobilized biomimetic or biospecific ligands and separates proteins according to principles of molecular recognition. HP-AC can be used for initial capture of proteins, as an intermediate step in a multiple step purification procedure, or for the affinity removal of unwanted high abundance proteins, provided a suitable affinity ligand for the target protein is available. HP-AC is highly selective and usually has a high capacity for the protein of interest. In HP-AC, analytes are eluted by step gradient or gradient elution, where the capture (loading) mobile phase is aqueous and of low ionic strength and the eluting mobile phase is aqueous and of higher ionic strength or of different pH value, or, alternatively, contains a mobile phase additive that competes with the target compound for binding to the immobilized biospecific ligand. In terms of achieving maximal selectivity and highest affinity in the interaction between the target substance(s) and the chromatographic sorbent, HP-AC excels over all other modes but each affinity sorbent must be tailored to the specific target compound.

Immobilized metal-chelate affinity chromatography (IMAC) exploits the affinities of the side chain moieties of specific surface-accessible amino acids in peptides and proteins for the coordination sites of immobilized transition metal ions. Most investigations employed tri- or tetradentate ligands, such as iminodiacetic acid (IDA), nitrilotriacetic acid (NTA), tris-(carboxy methyl)ethylene-diamine (TED), O-phosphoserine (OPS), or carboxymethylaspartic acid (CMA). The retaining mobile phases are aqueous with neutral pH and high ionic strength, the eluting mobile phases are of low pH and contain competing ligands or EDTA. Novel immobilized chelate systems, such as 1,4,7-triazacyclononane (TACN), however, show different chromatographic properties compared to the IMAC behavior of traditional chelating ligands. These AC techniques, in conjunction with soft gel matrices, have been applied in diverse analytical and preparative protein purifications. HP-AC separations can be performed with immobilized chemical or biological ligands, or with molecular imprinted polymers (MIPs).

Novel procedures to immobilize an IMAC ligand at the surface of silica supports have provided guidelines for the design of very stable HP-IMAC systems for peptides and proteins.

Table 10.7 summarizes the nature of the retaining and eluting mobile phases for the above-described chromatographic modes. Guidelines on how to choose a mode or a combination of modes for a specific separation task will be given in Section 10.5.

Table 10.7 Chromatographic mode employed in peptide and protein separation and their stationary and mobile phase characteristics.

Chromatographic mode	Stationary phase	Retaining mobile phase	Eluting mobile phase
SEC or GPC	Porous	Non retentive	Aqueous, low salt
RPC	Hydrophobic	Aqueous	Organic solvents
NPC	Polar	Nonpolar organic solvents	Polar organic solvents
HILIC	Hydrophilic	Nonpolar organic solvents	Polar organic solvents, water
ANPC	Polar	Organic	Aqueous
HIC	Mildly hydrophobic	Aqueous, high ionic strength	Aqueous, low ionic strength
AEX	Charged	Aqueous, high pH, low ionic strength	Aqueous, high pH, high ionic strength (or low pH), high selectivity counter-ion
CEX	Charged	Aqueous, low pH, low ionic strength	Aqueous, low pH, high ionic strength (or high pH)
AC	Biomimetic, biospecific	Low ionic strength	High ionic strength, competing ligand
IMAC	Metal chelate	Aqueous, neutral pH, high ionic strength	Low pH, competing ligand, EDTA

10.5 Method Development from Analytical to Preparative Scale Illustrated for HP-RPC

HP-RPC is currently the most frequently used high-performance liquid chromatographic mode for the analysis and preparative purification of peptides and proteins, in particular for applications that involve off-line or on-line electrospray ionization (ESI) mass spectrometry. The development of a method for preparative HP-RPC purification for the purpose of isolation of one or more component(s) from a peptide or protein product sample (or alternatively the purification of a synthesized product) is usually performed in four steps:

1. development, optimization, and validation of an analytical method;
2. scaling up of this method to a preparative chromatographic system;
3. application of the preparative method to the fractionation of the product;
4. finally, analysis of the individual fractions.

10.5.1 Development of an Analytical Method

The development of an analytical method for the separation of a peptide or protein encompasses the selection of the stationary and mobile phase taking into consideration the analyte properties (hydrophobicity/hydrophilicity, acid–base properties, charge, temperature stability, molecular size). This is followed by a systematic optimization of the (isocratic or gradient) separations, using either aliquots of the crude extract or, if available, analytical standards.

In the selection of the stationary and mobile phase, various chemical and physical factors of the chromatographic system that may contribute to a variation in the resolution and recovery of peptides and proteins need to be considered. The stationary phase contributions relate to the particle size, particle size distribution, particle compressibility, surface area, pore diameter, pore diameter distribution, ligand composition, ligand density, and surface heterogeneity. Typically, a particular HP-RPC material will be selected empirically as the starting point for the separation, taking into consideration its suitability for the separation task at hand, published procedures for similar types of peptides/proteins, availability of the stationary phase material for preparative chromatography, and whether information is available on the analyte properties. The mobile phase contributions relate to the type of organic solvents, eluent composition, ionic strength, pH, temperature, loading concentration, and volume.

Since the quality of a separation is determined by the resolution of individual peak zones, method development always aims to optimize the resolution. The method development for analytical separations always focuses on the least well-resolved peak pair(s) of interest, the so-called "critical" peak pair. The resolution depends on the column efficiency or plate number N, the selectivity α, and the retention factor k, all of which can be experimentally influenced through systematic changes in individual chromatographic parameters. In the isocratic mode of separation, resolution is determined from:

$$R_S = \frac{1}{4} N^{\frac{1}{2}} (\alpha - 1) \left(\frac{k}{1+k} \right) \tag{10.12}$$

As detailed above, the plate number N is the efficiency of the column and is a measure of the column performance. The selectivity α describes the selectivity of a chromatographic system for a defined peak pair and is the ratio of the k-values of the second peak to the first peak. The retention factor k is a dimensionless parameter and is defined as $k = (t_R - t_0)/t_0$ where t_R is the retention time of a particular peak and t_0 is the column void time. In this manner, normalization of the relative retention can be achieved for columns of different dimensions. Whilst N and α change only slightly during solute migration through the column, the value of k can be readily manipulated through changes in the elutropicity of the mobile phase by a factor of 10 or more. The best chromatographic separations for low or mid molecular weight analytes are generally achieved with mobile phase–stationary phase combinations that result in k values between 1 and 20.

In gradient elution, in contrast to isocratic elution, $\overline{N}$, $\overline{\alpha}$, and $\overline{k}$ are the median values for N, α, and k, since they change during the separation as the shape and duration of the gradient changes.

The "gradient" plate number $\overline{N}$ has no influence on the selectivity or the retention (except for temperature change). The selectivity $\overline{\alpha}$ and the retention factor $\overline{k}$ usually have only a minor influence on $\overline{N}$. While $\overline{N}$ and $\overline{\alpha}$ change only slightly during the solute migration through the column, the $\overline{k}$-value can change by a factor of 10 or more depending on the gradient steepness. Again, the best chromatographic separation is generally achieved with a value between 1 and 20. Although resolution in isocratic and gradient elution is mainly influenced by the mobile phase variables α or $\overline{\alpha}$) and k (or $\overline{k}$) and is nearly independent of N (or $\overline{N}$), for a given column, an optimization strategy should nevertheless start with appropriate selection of the stationary phase. This is because many initial choices (like column dimensions, choice of ligand, etc.) are determined by the overall strategy (i.e., separation optimization for quantification of several analytes or separation optimization for planned *scaling-up* to preparative purification of specific target compounds) and by the purification goals. Several computer-assisted, expert systems are commercially available to guide this selection. After the column is selected, with consideration of the above equation, the separation optimization is performed in three steps: optimization of the column efficiency N, then optimization of the selectivity α, and finally optimization of the retention factor $\overline{k}$-values.

Optimization of the Column Efficiency Optimization of the peak efficiency, expressed as the theoretical plate number, N, requires an independent optimization of each of the contributing factors that influence the band-broadening of the peak zones due to column and the extra-column effects. With a particular sorbent (ligand type, particle size, and pore size) and column configuration, this can be achieved through optimization of linear velocity (flow rate), the temperature, detector time constant, column packing characteristics, and by minimizing extra-column effects, by, for example, using zero-dead volume tubing and connectors. The flow rate (or alternatively the linear flow velocity) to achieve the minimum plate height, H, for a particular column can be taken from the literature or experimentally determined according to published procedures. The temperature of the column and the eluents should be thermostatically controlled to facilitate the reproducible determination of the various column parameters and to ensure resolution reproducibility.

Optimization of the Selectivity A change in selectivity α of the separation is the most effective way to influence resolution. This is mainly achieved by changing the chemical nature or concentration of the organic solvent modifier (acetonitrile, methanol, ethanol, isopropanol, etc.) in conjunction with the appropriate choice of mobile phase additive(s). However, if different organic solvents are used, different eluotropic strengths must be considered in order to allow elution of the analytes of the sample in the appropriate retention factor range. The interconversion of isocratic data to gradient data and vice versa can be achieved through the use of algorithms based on linear and nonlinear solvent strength theory.

Optimization of the Retention Factors Further optimization should focus on achieving the most appropriate retention factor for the different peptides or proteins in the mixture. In the isocratic elution mode of HP-RPC, resolution optimization can take advantage of the relationship between the retention time of an analyte (expressed as the retention factor k) and the volume fraction of the organic solvent modifier, φ. Although typically these dependencies are curvilinear, for practical convenience they are often treated as linear relationships. Thus, the change in retention factor as a function of φ can be represented by:

$$\ln k = \ln k_0 - S\varphi \tag{10.15}$$

where k_o is the retention factor of the solute in the absence of the organic solvent modifier and S is the slope of the plot of ln k versus φ. The values of ln k and S can be calculated by linear regression analysis. Greater precision in the quality of fit of the experimental data, and thus improved reliability in the prediction of the retention behavior of analytes in HP-RPC systems for mobile phases of different solvent composition, can be achieved through the use of an expanded form of Equation 10.15:

$$\ln k = \ln k_0 - S\varphi + S'\varphi^2 - S''\varphi^3 + \ldots \tag{10.16}$$

Similarly, in gradient elution HP-RPC, resolution optimization can take advantage of the relationship between the gradient retention time of an analyte (expressed as the median

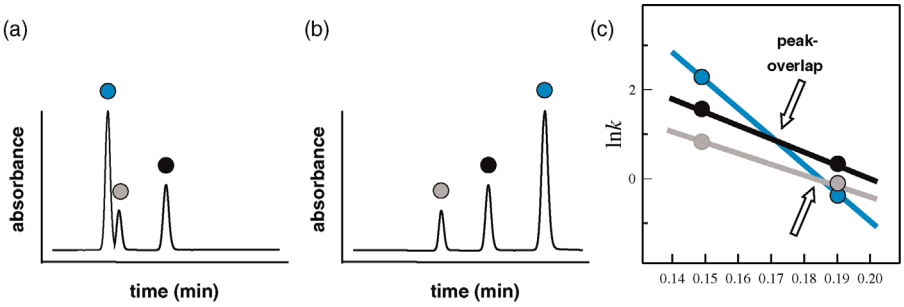


Figure 10.3 Optimization of isocratic elution. Two chromatograms obtained for (a) 19% and (b) 14% (v/v) of organic solvent modifier in the mobile phase (corresponding to $\varphi = 0.19$ and 0.14, respectively) can be used to plot (c) the corresponding logarithmic retention factors ln k versus the volume fraction of the organic solvent modifier to identify the mobile phase composition resulting in optimal peak spacing. Trend lines can determine mobile phase compositions that result in peak overlap or excellent peak resolution.

retention factor $\overline{k}$) and the median volume fraction of the organic solvent modifier, $\overline{\phi}$ in regular HP-RPC systems based on the concepts of the linear-solvent-strength theory, such that:

$$\ln \overline{k} = \ln k_0 - S\overline{\phi} \tag{10.17}$$

A mapping of the dependence of analyte retention (expressed as the natural logarithm of the retention factor k) on the mobile phase composition (expressed as the volume fraction of solvent in the mobile phase, φ) in isocratic elution (or as $\overline{k}$ versus $\overline{\phi}$ in gradient elution), with a minimum of two initial experiments, can be used to define the useful range of mobile phase conditions, and can indicate the mobile phase composition at which the band spacing is optimal (Figure 10.3).

Irrespective of whether the data are obtained through isocratic or gradient elution techniques employing two initial experiments (differing only by their mobile phase composition or gradient run times, respectively) with tracking and assignment of the peaks, a relative resolution map (RRM) can be established, which plots resolution R_S against the separation time (or gradient run time t_G). In the case of gradient elution, the RRM then allows determination of the optimal gradient run time (and gradient range). Such a procedure can be performed in any laboratory using Excel spreadsheets or through software packages (e.g., DryLab, LabExpert, etc.). Such strategies greatly reduce the time needed to achieve an optimal separation, as well as saving on solvent, reagent, and analyte consumption, and if full exploited permit instrumentation to be operated in a nearly fully automated, unattended fashion.

Optimization can also be performed via computer simulation software (e.g., Simplex methods, multivariate factor analysis programs, DryLab G/plus, LabExpert, etc.). In such procedures, resolution of peak zones is optimized through systematic adjustment of mobile phase composition by successive change in the φ-value (or equivalent parameters, such as the concentration of the ion pairing reagent employed).

In gradient elution, advantage is taken of a strategy with the following eight steps:

1. performance of initial experiments;
2. peak tracking and assignment of the peaks;
3. calculation of ln k_0 and S-values from initial chromatograms;
4. optimization of gradient run time t_G over the whole gradient range;
5. determination of new gradient range;
6. calculation of new gradient retention times t_g;
7. change of gradient shape (optional);
8. finally, verification of results.

Examples where such systematic method development has been used for the analytical separation of peptides and proteins can be found in the literature.

10.5.2 Scaling up to Preparative Chromatography

While analytical HPLC aims at the quantification and/or identification of compounds (with the sample going from the detector to waste), preparative chromatography aims at the isolation of compounds (with the sample going to the fraction collector). For preparative separations, method development always focuses on the peaks of interest, and the two adjacent eluting peaks. Optimization of the resolution of the peak of interest from the adjacent peaks has to take into account the sample size and the relative abundances of the components that form the basis of the separation task.

Table 10.8 Operating ranges of HPLC column types.

Column type	Sample quantity range	Column diameter (mm)	Column lengths (mm)	Flow rate range (ml min^{-1})
Preparative LC	mg–g	>4	15–250	5–20
Analytical LC	µg–mg	2–4	15–250	0.2–1
Micro LC	µg	1	35–250	0.05–0.1
Nano LC	ng–µg	<1	50–150	<0.05
Nano chip	ng	<0.1	50	<0.01

Once an analytical method is established, it can be scaled up to a preparative separation by taking into consideration the operating ranges of the column (Table 10.8) or by deliberate column overloading.

The concept of scaling up or down implies that the performance features, for example, selectivity behavior, of the stationary phase material used for the analytical and the preparative separation is identical, with the exception of particle size. Experimental studies have established the fundamentals of scaling up and allowed the development of experimental methods for their validation. To obtain an equivalent elution profile, the flow rate needs to be adjusted for columns with different internal diameter, according to the following formula:

$$F_{\text{preparative}} = \left[\frac{r_{\text{preparative}}}{r_{\text{analytical}}}\right]^2 \times F_{\text{analytical}} \tag{10.18}$$

where F is the flow rate and r is the column radius of the preparative or analytical column.

Estimates of the loading capacity of a particular column material can usually be obtained from the manufacturer. The mass loadability for a scaled up separation can be calculated with the following formula:

$$M_{\text{preparative}} = \left[\frac{r_{\text{preparative}}}{r_{\text{analytical}}}\right]^2 \times M_{\text{analytical}} \times C_{\text{L}} \tag{10.19}$$

where M is the mass, r is column radius of the preparative or analytical column, and C_{L} is the column length ratio.

In some cases, despite some loss of resolution, column overloading is an economic and viable method for compound purification. In analytical liquid chromatography, the ideal peak shape is a Gaussian curve. If under analytical conditions a higher amount of sample is injected, peak height and area change, but not peak shape or the retention factor. However, when more than the recommended amount of sample is injected onto the column the adsorption isotherm becomes nonlinear. As a direct consequence resolution decreases, and peak retention times and peak shapes may change.

There are two methods of column overloading, volume overloading and concentration overloading. In volume overloading, the concentration of the sample is maintained, but the sample volume is increased. The retention factor of the compound(s) increase(s), with broadened peak shape. In concentration overloading the volume of the injected sample is maintained, whilst the sample concentration is increased. The retention factors of the compound(s) decrease, and the peak may become *fronting* or *tailing* with triangular peak shape. The applicability of this method is limited by the solubility of the target compound(s) in the mobile phases employed. In preparative HPLC volume overloading is the method of choice, since it allows the separation of larger sample amounts than with concentration overloading. However, in praxis, a combination of both methods is usually applied.

10.5.3 Fractionation

Fraction collection can be manual, with a manually pressed button to start and stop collection, time based, with a fraction collecting during fixed pre-programmed time intervals, peak based, based on a chosen threshold of the up- and down-slope of a detector signal, and mass based, with fraction collection occurring only if the specific mass of a trigger ion is detected by mass spectrometry. Whatever the type of fraction collection, attention has to be given to the fraction collection delay times and a delay time measurement needs to be performed. For a peak with start time t_0 and end time t_{E}, fraction collection needs to be started when the start of the peak

arrives at the diverter valve ($t_0 + t_{D1}$) and ended when the end of the peak arrives at the needle tip ($t_E + t_{D1} + t_{D2}$), where t_{D1} is the delay time between detector and valve and t_{D2} is the delay time between valve and needle tip. In addition, a recovery collection can be performed, by which everything that is not collected as a fraction goes into a dedicated container.

10.5.4 Analysis of Fractionations

After the fractions have been collected, the solvent needs to be removed either by using a freeze dryer, rotary evaporator, or high-throughput parallel evaporator. Non-volatile components can be removed with reversed-phase SPE procedures prior to solvent removal if the aqueous portion of the buffer is sufficiently large. In the absence of on-line mass spectrometry, fractionation is usually accompanied by an off-line mode of quality analysis, consisting of a pre-preparative analysis of the unpurified material and a post-preparative analysis of the individual fractions. The pre- and post-preparative analysis can performed with analytical HPLC, MS, and activity testing of the biological compounds, if an assay is available.

10.6 Multidimensional HPLC

Although HPLC is a powerful separation technique, very often more than one chromatographic step is necessary to achieve a required degree of purity of the target compounds. In practice, this is achieved through a series of purification steps. As there are material losses associated with each purification step in these procedures the overall recovery of the product has to be optimized. This can be achieved by minimizing the number of purification steps employed. Multidimensional (multistage, multicolumn) high-performance liquid chromatography (MD-HPLC) offers the possibility of cutting the elution profiles into consecutive fractions, whereby these fractions can be treated independently. One important consequence of this strategy is the gain in peak capacity, defined as the number of peaks that can be accommodated between the first and the last peak in a separation of defined resolution. MD-HPLC has the potential of independent optimization of the separation conditions for each fraction and allows a relative enrichment or depletion of components. MD-HPLC can be applied to the purification of a particular peptide or comprehensive fractionation of complex peptide and protein mixtures.

10.6.1 Purification of Peptides and Proteins by MD-HPLC Methods

Protein Purification, Chapter 1

Peptides and proteins are generally isolated from complex biological matrices, starting with fractionation of the crude extract. Here, the chemical and physical properties of the target compound and those of the matrix determine the choice of the separation methods. In some cases this is preceded by a protein precipitation with salt or organic solvents. Then the crude extract is clarified (i.e., by filtration, centrifugation) to free the sample from particulate matter and make it suitable for chromatography. For the following separation step an appropriate sample buffer compatible with the mobile phase(s) of the particular chromatographic mode used in the next step is chosen. This is followed by an enrichment step, preferably using solid-phase extraction (SPE) or restricted access materials (RAMs) in a step elution mode to eliminate the majority of low molecular materials and to drastically reduce the volume of the sample. The next step consists of the intermediate purification of target compound(s) and a final chromatographic purification using various high-performance chromatographic modes of different selectivity (i.e., separating the analytes according to their molecular size, hydrophobicity/hydrophilicity, charge, biospecificity, etc.) according to the principles of chromatography summarized in Table 10.9.

Table 10.9 Optimization priorities at individual stages for a multidimensional, three-stage purification process after initial sample extraction exemplified for peptides and proteins and suitable chromatographic modes.

Purification stage	High priority	Lower priority	Chromatographic mode
Enrichment	High speed, high capacity	Resolution	AC, IMAC, IEX, HIC
Intermediate purification	High capacity, high resolution	Speed, recovery	IEX, HIC, SEC
Final purification	High resolution, high recovery	Speed, capacity	RPC, SEC

To purify a particular peptide or protein, it is often possible to select complementary chromatographic modes that allow the target compound to be obtained in high purity with only a few separation steps, preferably three steps or less. In such non-comprehensive MD-HPLC, only some of the analytes (as a single fraction) eluting from the first column are transferred to a second column for further purification (conventionally expressed by a hyphen, i.e., IEC-RPC). Such techniques are fast but not comprehensive, since most of analytes are not subjected to separation in a second dimension. The main advantage of the technique is the improved resolution of compounds that co-elute in the first dimension. A key requirement for such a purification scheme is that subsequent stages of the separation are orthogonal, that is, the two separation modes are not correlated to each other in relation to their retention characteristics (i.e., selectivity). For a single chromatographic dimension, the partly contradictory objectives of speed, resolution, capacity, and recovery can usually not be maximized simultaneously (i.e., a high resolution can be achieved but at the expense of speed; a high speed separation can reduce resolution, etc.) (Figure 10.4).

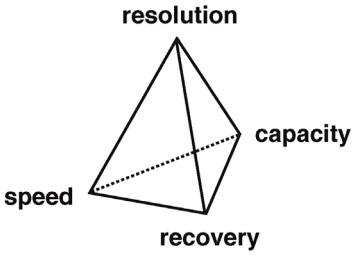


Figure 10.4 Optimization goals (speed, resolution, capacity, and recovery) for a chromatographic purification and their interrelationship.

A three-stage MD-HPLC protein purification process allows the overall purification objectives to be met by placing the emphasis for each purification stage on a different pair of objectives and choosing a chromatographic mode that is particularly well suited for the task and in a sequence that avoids time-consuming buffer exchanges (Table 10.9).

At the enrichment or capture stage, the emphasis is on speed and capacity, by employing HP-AC, HP-IEX, or HP-HIC, possibly in a solid-phase extraction format as a low resolution step. This stage aims at the initial isolation of the target product from the crude sample, at increasing its concentration, and also at the removal of major or critical contaminants.

At the intermediate purification stage, emphasis is placed on capacity and resolution, employing chromatographic modes with intermediate resolution, i.e., HP-AC, HP-IEX, HP-HIC, or HP-SEC. This stage has the objectives of removing the majority of impurities, that is, other proteins, nucleic acids, viruses, and endotoxins.

At the final chromatographic polishing step, emphasis is placed on resolution and recovery, employing high-resolution modes, such as HP-RPC. This stage strives to remove trace amounts of impurities or closely related compounds to obtain the final pure product. Therefore, HP-SEC may also be used as final polishing step, that is, to remove unwanted multimeric forms of the target protein.

In some cases, the capture and intermediate purification, or even the intermediate purification and final polishing step, may be achievable with a single separation step resulting in a two-stage purification process. In other cases, for example, for the purification of therapeutic proteins, four or more stages may be required to achieve the desired degree of protein purity.

10.6.2 Fractionation of Complex Peptide and Protein Mixtures by MD-HPLC

If the objective of a purification scheme is the comprehensive fractionation of a complex, multicomponent peptide mixture, as it is required in proteomics, it is of advantage to use orthogonal chromatographic modes, but such extensive fraction collection requires additional, sometimes substantial, infrastructure (e.g., a second HPLC pump, a thermostated autosampler, a robotic autosampler loader, switching valves, thermostated fraction collector and high-throughput evaporator, etc.). In comprehensive MD-HPLC, the entire analyte pool of the first column is transferred to the second column (expressed by a cross, i.e., IEC×RPC) as sequential aliquots, either successively onto one column or alternating onto two parallel columns. The resulting data can be represented as 3D contour plots, with retention times of the second dimension plotted against retention times of the first dimension. The information content of such comprehensive two-dimensional chromatograms is higher than that of individual one-dimensional chromatograms.

10.6.3 Strategies for MD-HPLC Methods

Regardless of which operational mode, off-line or on-line, is used, the compatibility of the mobile phases between successively employed chromatographic modes in a separation scheme needs to be considered. As a consequence, it may be necessary to process the fractions between two separation stages (e.g., through buffer exchange, concentration, or dilution) to enhance the compatibility of eluent composition of fractions from the first chromatographic dimension with the retaining mobile phase of the second chromatographic dimension. If a non-retentive chromatographic mode such as

size exclusion chromatography (SEC) is employed in conjunction with a retentive chromatographic mode, such as reversed-phase chromatography (RPC) or ion exchange chromatography (IEX), it is usually performed first. This allows relatively large eluent volumes stemming from isocratic elution in the non-retentive mode to be reduced through the capture of analytes under the retaining mobile phase conditions of the subsequent retentive chromatographic mode and reduction of extra-column band broadening with resulting loss of resolution.

In the *off-line* mode, the eluent of the first column is collected as fractions, either manually or with an automated fraction collector, and re-injected onto the second column. Typical processing steps may include dilution, volume reduction by freeze-drying or automated high-throughput parallel evaporation systems taking into account the boiling point(s) or volatility of the target analyte(s) and organic solvent if these are contained within the eluates. The use of volatile mobile phase additives then allows a relative fast buffer exchange.

The *on-line* mode uses high-pressure, multi-position, multi-port switching valves that allow selection of pathways for single fractions from the first chromatographic dimension to subsequent column(s) of the second chromatographic dimension. The fractions from the first dimension are transferred either directly or through one (or more) intermediate trapping columns for the purpose of concentration and automated buffer exchange. This approach requires complex instrumentation, resulting in increased optimization time and reduced system flexibility. It has, however, numerous advantages in terms of reproducibility, recovery, speed, and automation.

10.6.4 Design of an Effective MD-HPLC Scheme

MD-HPLC for peptides and proteins requires thoughtful selection of orthogonal and complementary separation modes, of the order of their utilization, and independent optimization with respect to the chromatographic goals (speed, resolution, capacity, and recovery). Furthermore, besides the mobile phase composition of the employed chromatographic modes, the elution mode (isocratic, step, or gradient elution) and flow rates and mobile phase temperatures need to be considered.

To exploit the full peak capacity of a two-dimensional system, it is advantageous if the applied chromatographic modes are orthogonal. It is generally accepted that the dimensions in a two-dimensional separation system are orthogonal, if the separation mechanism of the two dimensions are independent of each other. The codicil is that the distribution coefficients of analytes in the first dimension do not correlate with the distribution coefficients in the second dimension. An example of such orthogonality of different separation modes in high-performance liquid chromatography is ion-exchange chromatography (CEX or AEX) and reversed-phase chromatography (RPC) as they separate according to net charge or hydrophobicity, respectively. A very coarse classification of chromatographic modes commonly applied in the MD-HPLC of peptides and proteins according to their separation principles is depicted in Figure 10.5.

For an ideal orthogonal, two-dimensional separation the overall peak capacity (PC) is defined as the product of the peak capacities in each dimension:

$$PC_{2D\text{-}System} = PC_{First\ Dimension} \times PC_{Second\ Dimension} \quad (10.20)$$

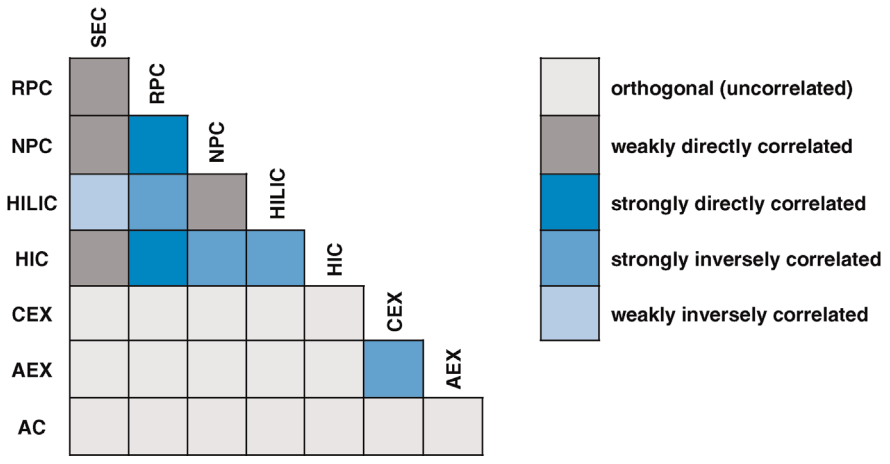


Figure 10.5 Degree of orthogonality of major chromatographic modes employed in the separation of peptides and proteins.

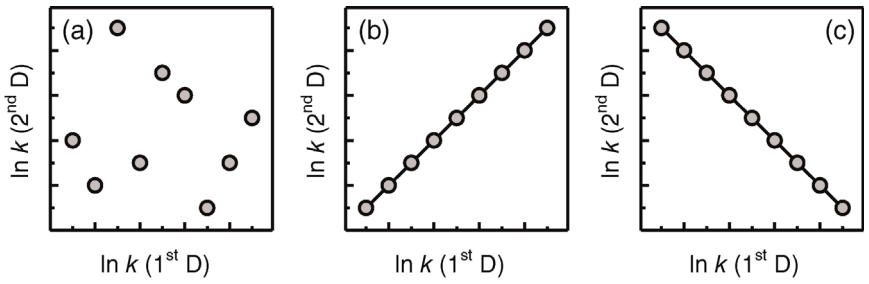


Figure 10.6 Two-dimensional separation space for a set of peptides and proteins (circles) utilizing separation systems that are (a) uncorrelated (orthogonal), (b) completely correlated, and (c) inversely correlated, where the retention factors obtained in the second dimension are plotted versus the retention factors obtained in the first dimension.

However, if two non-identical chromatographic modes with some degree of similarity are used in a 2D system, the increase in the peak capacity and the total number of analytes that can be separated is much lower than the product of the peak capacities of individual dimensions. The peak capacity also depends on the elution mode. Gradient elution provides a higher peak capacity than isocratic elution and is of advantage in 2D LC. Notably, since selectivity in chromatography depends not only on the stationary phase but on the mobile phase as well, orthogonal separations can be achieved through fine-tuning of the separation conditions, even if the principal separation mechanisms of both dimensions are similar. In addition, the structure of analytes has an effect on the peak capacity. In many separation systems, the contribution of structural units, especially the repeating units, to the Gibbs free energy of association of the analytes with the immobilized chromatographic ligands is additive. Such structural repeating units can be hydrophobic or polar. If one chromatographic system in a 2D LC has no selectivity for a structural element, then the first and the second dimension are non-correlated (orthogonal) with respect to the repeating structural unit (Figure 10.6a). In a completely correlated separation system, with correlated retention factors in the two dimensions, the separations space is not utilized and thus not ideal (Figure 10.6b). In inversely correlated 2D LC×LC separation systems, the retention time increases in the first dimension, but decreases in the second dimension (Figure 10.6c). Neither correlated or inversely correlated 2D LC×LC increase the peak capacity significantly. The peak capacity in 2D LC×LC decreases with increasing correlation of the selectivity between the first and the second chromatographic dimension. In practice however, 2D LC×LC systems are rarely fully orthogonal with respect to each structure distribution type (i.e., hydrophobic, polar). Many partially orthogonal systems use only part of the theoretically available two-dimensional separation space but can be evaluated using analytes differing in the numbers of hydrophobic or polar structural units or by quantitative structure–retention relationships (QSRRs). Orthogonal systems with non-correlated selectivities provide the highest peak capacity and therefore the highest number of resolved peaks.

Although the suitability of a chromatographic mode employed in 2D LC×LC separations depends on the selectivity of the employed stationary phase, the selection of the mobile phase for each chromatographic dimension is of fundamental importance, to achieve maximal utilization of the two-dimensional separation space. In contrast to off-line 2D LC procedures, where the collected fraction can be subjected to evaporation, dilution, or extraction, before injection onto the column of the second dimension, the compatibility of the mobile phases in on-line 2D LC×LC in terms of miscibility, solubility, viscosity, and eluotropic strength is much more important. The compatibility of commonly used mobile phase of various chromatographic modes used for the separation of peptides and proteins is depicted in Figure 10.7.

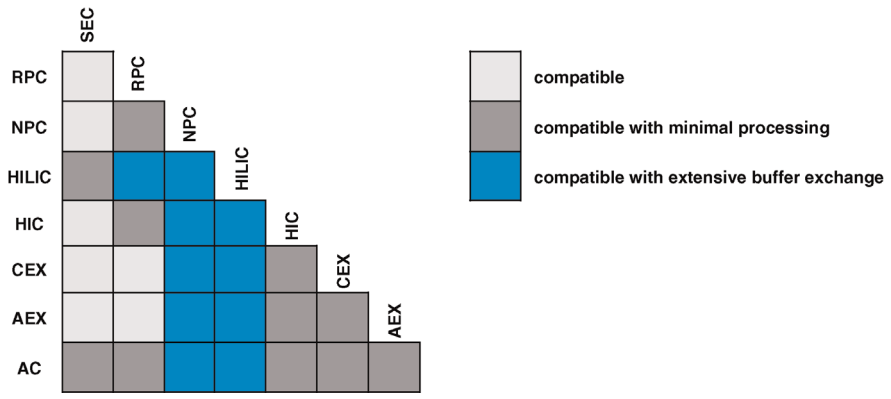


Figure 10.7 Compatibility between mobile phases of common chromatographic modes based on miscibility, solubility, and eluotropic strength.

10.7 Final Remarks

The comprehensive coverage of various high-performance liquid chromatographic (HPLC) methods for the analytical and preparative separation of peptides and proteins as attempted in this chapter may give reasons to believe that their possibilities for bioanalysis are already exhausted. However, there are numerous unresolved challenges in the field of proteomics, in process analytical technology (PAT), and medical diagnostics. The anticipated advances in the HPLC of peptides and proteins may thus lead to new discoveries in biology and medicine.

To this end, it will also be the responsibility of analytical chemists and biochemists to ensure that the development of new separation and preparative process methods occurs according to the principles of green analytical chemistry, considering the issues of waste minimization and hazard reduction. The first steps have already been made, but there is still a tremendous potential for investigators to pursue new and sustainable aspects of method development, a pursuit that hopefully has been encouraged by this chapter.

Further Reading

Anderson, J., Berthod, A., Pino, V. and Stalcup, A.M. (2016) *Analytical Separation Science*, Volumes 1-5, Wiley-VCH Verlag GmbH, Weinheim.

Carta, G. and Jungbauer, A. (2010) *Protein Chromatography: Process Development and Scale-Up*, Wiley-VCH Verlag GmbH, Weinheim.

Fanali, S., Haddad, P.R., Poole, C., Schoenmakers, P., and Riekkola, M.L. (eds) (2017) *Liquid Chromatography: Fundamentals and Instrumentation*, 2nd edn, Elsevier.

Fanali, S., Haddad, P.R., Poole, C., Schoenmakers, P., and Riekkola, M.L. (eds) (2017) *Liquid Chromatography: Applications*, 2nd edn, Elsevier.

Hearn, M.T.W. (1991) *HPLC of Proteins, Peptides and Polynucleotides*, VCH Verlag, Weinheim.

Henschen, A., Hupe, K.P., Lottspeich, F., and Voelter, W. (eds) (1985) *High-Performance Liquid Chromatography in Biochemistry*, VCH Verlag, Weinheim.

Kellner, R., Lottspeich, F., and Meyer, H.E. (eds) (1999) *Microcharacterization of Proteins*, 2nd edn, Wiley-VCH Verlag GmbH, Weinheim.

Mant, C.T. and Hodges, R.S. (1991) *High-Performance Liquid Chromatography of Peptides and Proteins: Separation, Analysis, and Conformation*, CRC Press, Boca Raton, FL.

Meyer, V. (2010) *Practical High-Performance Liquid Chromatography*, 5th edn, John Wiley & Sons Ltd, Chichester.

Lundanes, E., Reubsaet, L., and Greibrokk, T. (2013) *Chromatography: Basic Principles, Sample Preparations and Related Methods*, Wiley-VCH Verlag GmbH, Weinheim.

Schmidt-Traub, H., Schulte, M., and Seidel-Morgenstern, A. (eds) (2012) *Preparative Chromatography*, 2nd edn, Wiley-VCH Verlag GmbH, Weinheim.

Simpson, R.J. (2003) *Proteins and Proteomics: A Laboratory Manual*, Cold Spring Harbor Laboratory Press.

Simpson, R.J. (2003) *Purifying Proteins for Proteomics: A Laboratory Manual*, Cold Spring Harbor Laboratory Press.

Simpson, N.J.K. (2000) *Solid-Phase Extraction: Principles, Techniques, and Applications*, Marcel Dekker.

Simpson, R.J., Adams, P.D., and Golemis, E.A. (2008) *Basic Methods in Protein Purification and Analysis: A Laboratory Manual*, Cold Spring Harbor Laboratory Press.

Snyder, L.R., Kirkland, J.J., and Dolan, J.W. (2010) *Introduction to Modern Liquid Chromatography*, John Wiley & Sons, Inc., Hoboken.

Snyder, L.R., Kirkland, J.J., and Glajch, J.L. (1997) *Practical HPLC Method Development*, 2nd edn, John Wiley & Sons, Inc., New York.

Staby, A. and Rathore, A.S. (2017) *Preparative Chromatography for Separation of Proteins* (Wiley Series in Biotechnology and Bioengineering), Wiley-VCH Verlag GmbH, Weinheim.

Wixom, R.L. and Gehrke, C.W. (eds) (2010) *Chromatography – A Science of Discovery*, John Wiley & Sons, Inc., Hoboken.

Electrophoretic Techniques

11

Reiner Westermeier[1] and Angelika Görg[2]

[1]*85354 Freising, Germany*
[2]*Technische Universität München, TUM Emeriti of Excellence, Arcisstraße 21, 80333 München, Germany*

Electrophoresis is the migration of charged particles in an electric field. Differences in the charges and the sizes of particles cause different electrophoretic mobilities. Mixtures of substances are thereby separated into singular zones (Figure 11.1). Basically, three different electrophoretic methods are applied: zone electrophoresis (in a matrix or matrix-free) with a homogeneous buffer system, isotachophoresis with a discontinuous buffer system, and isoelectric focusing with a pH gradient.

All electrophoretic separation techniques afford a very high resolution. They have been established in a broad spectrum of analytical and preparative applications, particularly in biochemistry and molecular biology, in clinical and forensic medicine, as well as in the taxonomy of microorganisms, plants, and animals.

Electrophoretic separations are carried out in different media:

- In free solution: In open capillaries or thin buffer layers (free flow electrophoresis) the sample components are mainly separated on the basis of the charge differences. Owing to the development of Joule heating during electrophoresis thermal convection can disturb the separation.
- In stabilizing matrices: Membranes or gels (anti-convective media) counteract zone dispersions, which are caused by convection. In porous matrices the motion of particles is differently retarded depending on their size and, therefore, the resulting migration velocities are influenced by both size *and* charge.

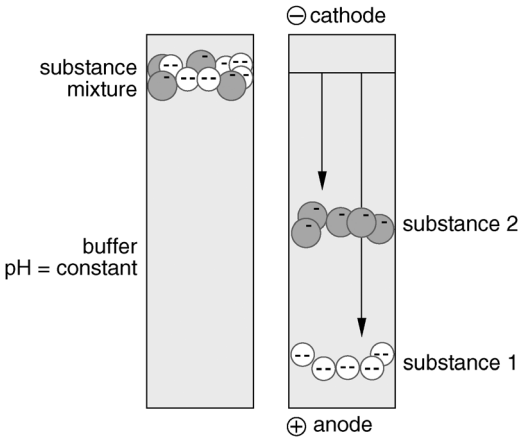


Figure 11.1 Separation principle of electrophoresis: electrically charged particles of different charges and sizes migrate in the electric field with different migration velocities; homologue substances form discrete zones.

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

Figure 11.2 Swedish postage stamp showing the first electrophoretic separation of human serum, which had been performed by Arne Tiselius.

11.1 Historical Review

The first electrophoresis method was developed in the 1930s by the Swedish scientist Arne Tiselius. He was awarded the Nobel Prize in Chemistry 1948 for his research on both electrophoresis and chromatographic adsorption analysis. Tiselius used a U-shaped tube, filled with buffer and with both limbs connected to an electric DC source to separate human serum into four main components: albumin and the α-, β-, and γ-globulins.

Figure 11.2 shows a typical result of the Tiselius method, visualized with Schlieren optics, displayed on a Swedish postage stamp. In combination with the results obtained by Theodor Svedberg with ultracentrifugation it could be proven that proteins are not differently assembled colloidal aggregates – as assumed until that date – but are macromolecules with defined sizes, shapes, and charges.

This first "moving boundary" electrophoresis technique was soon further developed to the method of zone electrophoresis: anti-convective media like paper, agar gels, and silica gels were introduced, onto which samples were applied in narrow zones. Owing to high intrinsic charges and the low sieving efficiency of these support media the resulting zones were blurred and poorly resolved. New matrices, though, were soon introduced: starch gels in 1955 by Smithies, cellulose acetate film by Kohn in 1957, polyacrylamide gels by Raymond and Weintraub in 1959, and agarose gels by Hjertén in 1961. Starch gels are still used for genetic studies and cellulose acetate film for clinical routine analysis. Agarose gels are mainly employed for the separation of DNA fragments and immune electrophoresis for specific and quantitative detection of proteins. Polyacrylamide gels are chemically inert, completely transparent, and exhibit the highest resolution for the separation of DNA fragment and proteins. Discontinuous polyacrylamide gel electrophoresis (disc PAGE), introduced by Ornstein and Davis in 1964, is the basis for numerous high-resolution electrophoresis techniques in gels and capillaries.

Thanks to the synthesis of carrier ampholytes, developed by Vesterberg in 1966, Svensson–Rilbe's concept of "natural" pH gradients was realized, leading to a new and very high resolving measure and separation principle for proteins, namely, isoelectric focusing. In the presence of an electric field the charged protein molecules migrate in a pH gradient until they reach the pH value corresponding to their isoelectric point, where the net charge of the molecule is zero. This means that its migration velocity at this point is also zero. In this way a separation is achieved and the isoelectric points can be determined in a simple way.

In 1967 two techniques for molecular size determination of proteins were launched: SDS polyacrylamide gel electrophoresis (SDS, sodium dodecyl sulfate) by Shapiro, Vinuela, and Maizel and porosity gradient gel electrophoresis by Margolis and Kenrick. In 1975 O'Farrell combined isoelectric focusing of denatured proteins with SDS PAGE to realize high-resolution two-dimensional electrophoresis, which for the first time allowed the separation of entire highly complex cell lysates and tissue extracts into all the protein constituents. By employing highly sensitive detection methods like autoradiography and fluorography several thousand protein spots can be visualized. In the same year Sanger developed DNA sequencing, which is also based on electrophoretic separation.

In addition, in 1975 the first blotting method was introduced by Southern. This involved the transfer of electrophoretically separated DNA fragments from agarose gels onto an immobilizing membrane with capillary forces, and subsequent hybridization. In 1979 this approach was modified for proteins by Towbin and colleagues: after electrophoretic transfer of the separated proteins they can either be immunologically identified (the so-called "Western blotting") or their amino acid composition and sequence are determined on the membrane.

The introduction of immobilized pH gradients in 1982 allowed a new concept of isoelectric focusing. In 1988 Görg established two-dimensional electrophoresis with immobilized pH gradients, resulting in improved reproducibility and higher loading capacity. The development of capillary electrophoresis in 1983 opened up new analytical possibilities. It will be described in the next section. Parallel with the analytical electrophoresis techniques, several preparative methods have been developed, like free flow electrophoresis and isoelectric focusing in dextran flatbeds, with columns filled with sucrose gradients, or between isoelectric membranes.

Capillary Electrophoresis, Chapter 12

11.2 Theoretical Fundamentals

Capillary Electrophoresis, Chapter 12

The electrophoretic mobility is a substance specific dimension, which determines the migration velocity in the electric field and is, therefore, essential for the separation.

Various forces affect a charged particle in an electric field, such as an accelerating force F_e, which acts on the charge q of the particle:

$$F_e = qE \text{ with } q = ze \tag{11.1}$$

and the friction force F_{fr}, which acts as a decelerating force:

$$F_{fr} = f_c v \tag{11.2}$$

where E is the electric field strength, v is the migration velocity of the particle, and f_c is the friction coefficient. The friction coefficient is dependent on the viscosity of the medium and, if applicable, the pore size of the matrix.

The equilibrium between these two forces causes the particle to move in the electric field with a constant velocity:

$$F_e = F_{fr}; \quad qE = f_c v \longrightarrow v = \frac{qE}{f_c} = uE \tag{11.3}$$

The proportionality factor between migration velocity and field strength is the substance specific dimension u, the mobility.

For small globular particles Stokes law can be applied to calculate the friction force, which leads to the following equation for the mobility:

$$u = \frac{q}{f_c} = \frac{ze}{6\pi\eta r} \tag{11.4}$$

where: z is the charge number,
e is the elementary charge in Coulomb,
η the viscosity of the solution,
r is the Stokes radius of the particle (i.e., the radius of the hydrated ion).

For non-globular particles like peptides and proteins an empirical correlation between the molecular mass M and the mobility can be indicated:

$$u = \frac{q}{M^{\frac{2}{3}}} \tag{11.5}$$

Alternative values for the exponent of the molecular mass between $\frac{1}{3}$ and $\frac{2}{3}$ can be found in the literature.

In infinitely diluted protein and peptide solutions two more forces become important that act on a charged particle, namely, the relaxation force and the retardation force, which are generated by the ionic atmosphere of the particle. According to the Debye–Hückel theory each particle is encircled by a counter-charged ionic atmosphere, whose radius β is dependent on the ionic strength. The force the electric field exerts on the ions of the ionic atmosphere is transferred onto the solvent molecules. Therefore, the central ion does not migrate through the stationary liquid but through a solution flowing in the opposite direction. This retardation effect causes a reduction of the central ion's velocity.

When the electric field is applied, this ion cloud "limps" behind the central ion and exerts an electric force that decelerates the central ion. This effect is called the relaxation effect.

Owing to these two effects the mobility decreases with increasing ionic strength. The different forces acting on a charged particle in an electric field and their contact points are displayed in Figure 11.3.

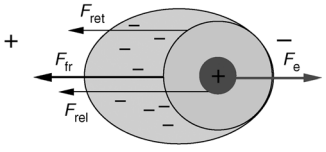


Figure 11.3 Accelerating and decelerating forces that act in an electric field on a charged, hydrated particle with an ion cloud (F_e = accelerating force, F_{fr} = friction force, F_{ret} = retardation force, F_{rel} = relaxation force).

For weak acids and bases it is not the mobility of the fully dissociated particles that determines the migration velocity, but the effective mobility u_{eff}, which is linked with the ion mobility via the degree of dissociation α (the relation between cation or anion and the total concentration of the electrolyte):

$$u_{eff} = \sum_i \alpha_i u_i \tag{11.6}$$

This means that for weak acids and bases (e.g., peptides and proteins) the migration velocity – and thereby the resolution – can be optimized via modification of the pH of the electrolyte.

The specific conductivity K of a solution results from the effective mobilities of all particles in a solution as follows:

$$K = F \sum_{i=1} c_i u_i |z_i| \tag{11.7}$$

where F is the Faraday constant and c is the concentration of the single ionic species. In electrophoresis generally buffer systems composed of an acid and a base are employed, for instance Tris-chloride, Tris-borate. In the electric field not only the sample ions are migrating, but also – and to a high extent – the ions of the dissociated buffer components. Thus, for electrophoresis, buffer reservoirs are required at both electrodes. The buffer concentrations and amounts in these reservoirs must be sufficiently high to prevent depletion of the buffer.

During an electrophoretic run Joule heat is generated by the electric current. For the heat W generated per volume unit the following equation applies:

$$W = E^2 \lambda c \tag{11.8}$$

where c is the molecular concentration of the electrolyte, λ is the equivalent conductivity, and E is the electric field strength. For the equivalent conductivity the following applies:

$$\lambda = u_i z_i F \tag{11.9}$$

Heat dissipation occurs through the walls or one of the lateral sides of the system. Thereby a temperature gradient develops, which can cause convective mixing in electrophoresis in free solution. To keep temperature differences as low as possible, narrow capillary diameters, with, respectively very thin layers or gels, as well as material with good temperature conductivity and thin walls should be employed. A further requirement to obtain the maximum separation effect is efficient temperature dissipation via liquid cooling.

The multitude of materials used like glass, *fused silica,* Teflon, paper, agarose, and cellulose acetate film carry surface charges, and generate an electrochemical double layer when they are in contact with an electrolyte solution. As an example, the structure of such a double layer is described here for capillaries made from fused silica, a material that has been intensively studied. The dissociation of the silanol groups (–SiOH) causes the formation of negative charges on the capillary wall. These negative charges are compensated by positive counter charges at the interface of the solution. This leads to a drop in potential as shown in Figure 11.4. The double layer consists of a rigid and a diffuse layer, whereby the potential drop is linear in the rigid layer and exponential in the

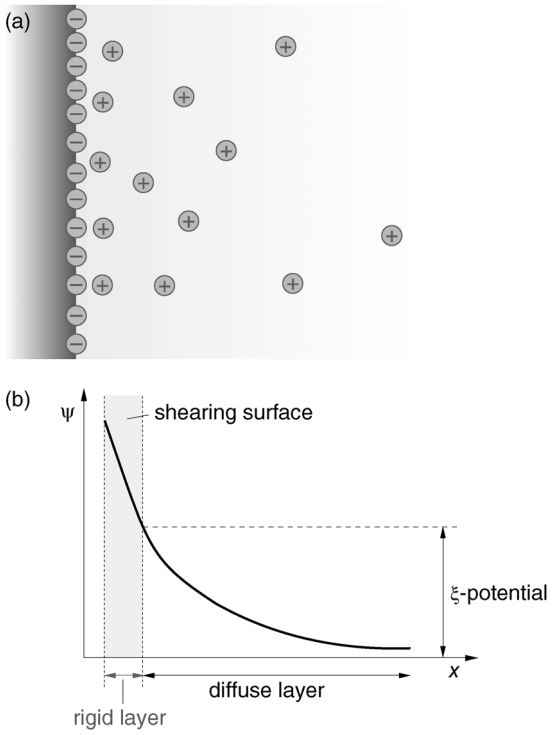


Figure 11.4 Structure (a) and potential (Ψ) gradient (b) of an electrochemical double layer; x is the distance from the capillary wall.

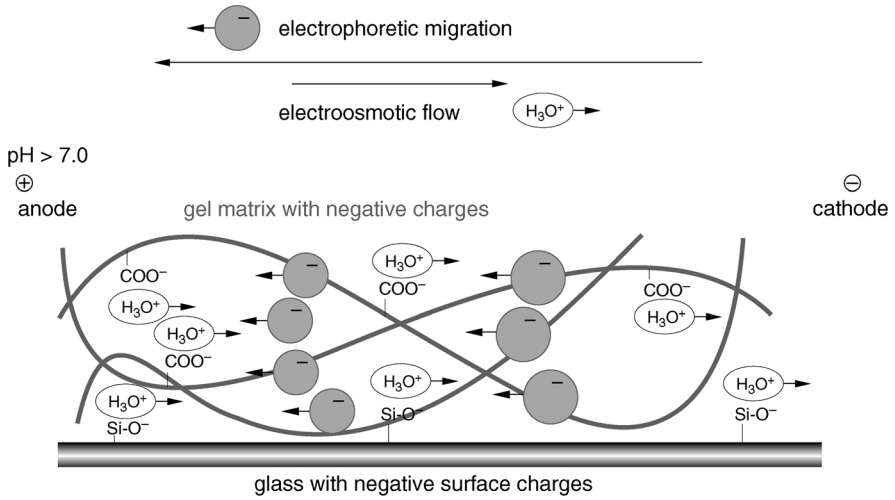


Figure 11.5 Electroendosmosis: At pH values above pH 7 silanol groups on glass surfaces and carboxylic and sulfonyl groups in gels become negatively charged. When the matrix and/or the surface of the equipment carries fixed charges, in the electric field an electric flow is generated, which is directed in the opposite direction to the electrophoretic migration.

diffuse layer. The lower the ionic strength of the solution, the further the diffuse double layer reaches inside the solution.

When the electric voltage is applied, the positive counter charges cause a flow of the solvent in the cathodal direction, originated by impulse transfer into the solvent – the electroosmotic flow EOF. The velocity of the EOF (v_{EOF}) is dependent on the so-called ξ-potential (zeta potential), the potential in the shearing surface (frequently equal to the boundary surface between rigid and diffuse double layer), the electric field strength, as well as the viscosity η, and the dielectric coefficient ε in the double layer:

$$v_{EOF} = \frac{\epsilon \xi E}{4\pi\eta} \tag{11.10}$$

Thus, during electrophoresis a flow of the liquid phase occurs. In gel electrophoresis this is called electroendosmosis (synonymous with electroosmosis). Analogously to the migration of charged particles in a liquid medium effected by the electric field, there is a movement of an ionic solution in the vicinity of charged surfaces or/and inside a gel matrix (Figure 11.5). Because the direction of the electroosmotic flow is opposite to that of the migration direction of the sample ions, the electroendosmosis causes unwanted distortion and diffusion of the zone. Therefore, charge-free materials and separation media are preferred.

Agarose is available in different qualities, which are defined by their values of electroendosmosis m_r: from 0.25 (highly charged) to "0.00" (almost free of electroendosmosis).The electroendosmosis value is measured using a nonionic dye, for example, dextran blue, which is applied on the gel before electrophoresis. Figure 11.6 shows a result of the separation of several protein mixtures in an agarose gel with a medium m_r value. Owing to the blurred zones the resolution and the detection sensitivity are poor.

The electroosmotic flow effect also disturbs isoelectric focusing with free carrier ampholytes (Section 11.3.12): because the EOF is pH dependent, different pH values in the system lead to additional diffusion mixing, which causes a drift of the pH gradient. In capillary electrophoresis, however, separations are frequently run with high EOF, because the electroosmotic flow profile is stamp-shaped and does not contribute to peak broadening in an open tube.

A further important dimension for electrophoretic separations is dispersion. To achieve a maximum resolution between the sample components it is important to keep the starting profile, which means the width of the applied sample zone, as narrow as possible and to prevent an excessive dispersion (broadening) of the starting zone during the separation process. Under ideal conditions there is only diffusion in the separation direction; in practice it is impossible to completely exclude additional effects due to temperature gradients, electroendosmosis, sample adsorption, and sample application profiles that are too wide.

The total peak broadening is described by the standard deviation of the concentration dispersion (σ_{tot}), which is composed of the individual contributions (diffusion, injection,

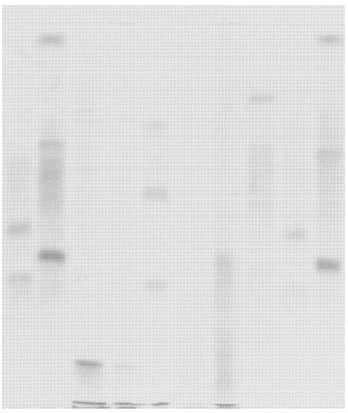


Figure 11.6 Protein electrophoresis in an agarose gel with medium electroendosmosis. The bands are blurred.

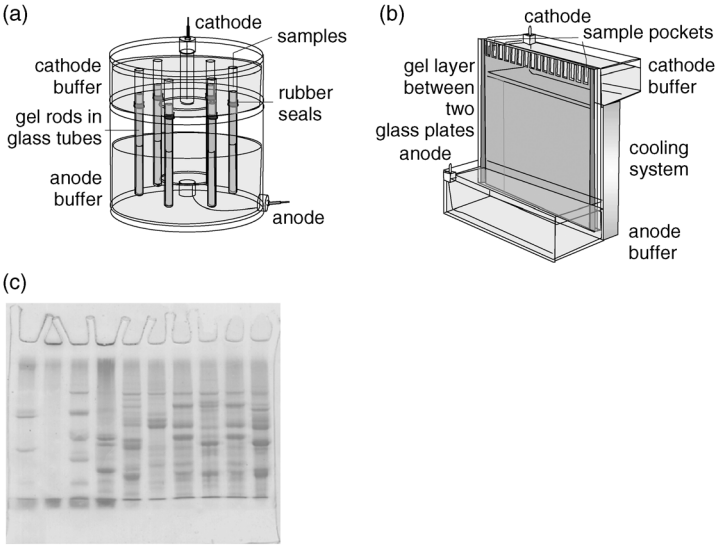


Figure 11.7 Vertical electrophoresis apparatuses for polyacrylamide gel electrophoresis and isoelectric focusing. The electrode wires are located in the buffer tanks. (a) Gel rod apparatus. (b) Slab gel apparatus. The sample wells are formed with a comb during gel casting. (c) Stained vertical gel after a protein separation.

temperature gradient, electroendosmosis, adsorption) according to the addition of the variances:

$$\sigma_{tot}^2 = \sigma_{Dif}^2 + \sigma_{Inj}^2 + \sigma_T^2 + \sigma_{EOF}^2 + \sigma_{Ads}^2 + \ldots \quad (11.11)$$

The contribution of the injection profile can be markedly reduced by employing a discontinuous buffer system, which produces an isotachophoretic sample concentration (Section 11.3.7). The temperature effect can be reduced by efficient cooling, and sample adsorption can be reduced by the use additives in the buffer.

11.3 Equipment and Procedures of Gel Electrophoreses

This section describes the commonly used equipment for gel electrophoresis techniques. Special appliances and apparatuses for preparative techniques, electroelution, two-dimensional electrophoresis, and free flow and capillary electrophoresis will be described in respective sections. Basically, three apparatuses are needed: power supply, electrophoresis chamber, and a thermostatic circulator (chiller).

The range of power supplies used for gel electrophoresis is from 100 V/400 mA to a maximum of 5000 V/150 mA depending on the method. Programmable power supplies are very useful, when a series of different voltage and current limitations have to be set. Modern power supplies have to comply with several safety requirements in order to avoid electric shock hazards.

During electrophoresis Joule heat is produced, which must be dissipated. Native proteins and complexes are sensitive to high temperatures; this requires separations under cooling conditions. Chambers connected to thermostatic circulators deliver better and more reproducible results than chambers without a cooling device. Running electrophoresis instead in a cold room is not a good alternative, because air has poor temperature conduction.

In addition, the use of tap water for cooling is not recommended, because it is inexact and wastes drinking water. Cooling of gels is either performed directly through cooling plates or indirectly via the anodal or the cathodal buffer.

Gel electrophoresis is either performed in vertical or in horizontal flatbed systems. In vertical chambers the gels are completely enclosed in tubes or cassettes and the buffer (Figure 11.7a and b). Platinum wires are used as electrodes, which are positioned in the buffer tanks. The samples are weighted with sucrose or glycerol to avoid mixing with buffer, and then applied on top of the gels either into the tubes or in gel pockets with a syringe or a micropipette. The sample pockets

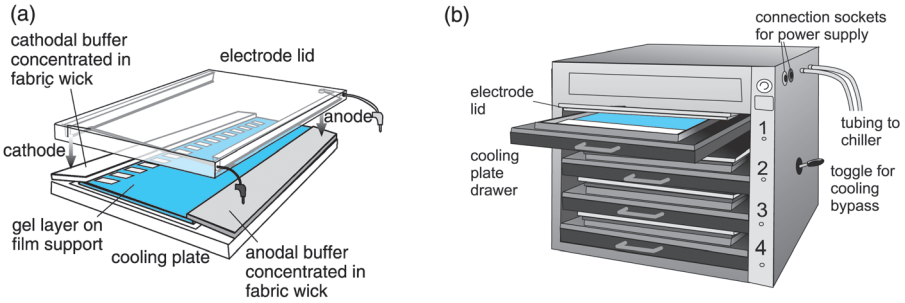


Figure 11.8 Horizontal electrophoresis system. Here cooling is very efficient. (a) Film-backed gels are mostly used. The platinum electrode wires are located on the lower side of the lid, which is laid directly onto the electrode wicks, which are soaked with concentrated buffer. (b) Electrophoresis tower with cooling plate drawers.

are molded with a comb, which is inserted into the cassette during gel casting and polymerization. Figure 11.7c shows a typical stained gel after vertical electrophoresis.

For horizontal systems film-backed gels with open surfaces are used (Figure 11.8a). The samples are directly pipetted into sample wells, which are molded with a template during gel casting, or they are applied with sample applicator strips or sample pieces made from cellulose and cotton. Because the gels are not hermetically enclosed during the separation, it is no problem to run gels of different formats in such an apparatus. Large buffer volumes and tanks are not needed, because fabric wicks soaked in concentrated buffers act as buffer reservoirs. The platinum electrode wires are placed directly onto the electrode wicks. The horizontal design has no problems with buffer leakage, and glass plates and spacers are not needed. Furthermore it is possible to apply thinner gel layers than in vertical systems. Thinner gels on film-backing are more efficiently cooled than vertical gels in cassettes, resulting in higher resolution. These features have a particularly positive impact on two-dimensional electrophoresis. Figure 11.8b shows an electrophoresis tower with cooling plate drawers for multiple flatbed electrophoresis separations.

For most gel electrophoresis separations voltages above 200 V are needed to achieve the necessary field strength. To avoid safety hazards in the laboratory the following safety advice here must be followed:

> Electrophoresis must be performed in closed apparatuses. Cables and plugs must be correctly insulated and correctly dimensioned. Power supplies must automatically switch off when a short circuit or ground leakage occurs. Electrophoresis equipment must be placed in a dry place. The electric connections of the separation chamber must be designed in such a way that the circuit is automatically disrupted upon accidentally opening a chamber. Electrophoresis systems must be set up so that potentially leaking buffer cannot get into the power supply.

For visualization of fluorescent labeled or stained proteins and the image analysis of electropherograms in gels and on membranes scanners, densitometers or camera systems are employed, which convert the results into digital signals. The qualitative and quantitative evaluation of the separations is carried out with a personal computer and the appropriate software; the results are stored in databases.

11.3.1 Sample Preparation

The protein solutions must not contain any solid particles or grease, because they would block the pores of the matrix and disrupt the separation. Therefore, the protein solutions should be degreased, filtered, and/or centrifuged prior to electrophoresis. Some proteins and enzymes are sensitive towards certain pH values or buffer components; this can cause configuration changes, denaturing, generation of complexes, and intermolecular interactions. Electrophoresis techniques are sensitive to high salt and buffer concentrations in the samples, as this will introduce additional ions into the system. The salt concentration should not exceed 50 mmol l^{-1}. To increase the solubility of proteins nonionic chaotropes like urea are added in high concentrations to the samples as well as to the gel. The solubility of hydrophobic proteins is further increased by adding nonionic detergents (e.g., Triton X-100) or zwitterionic detergents (e.g., CHAPS, ASB C7BzO, or other amidosulfobetaines).

Protein Purification, Chapter 1

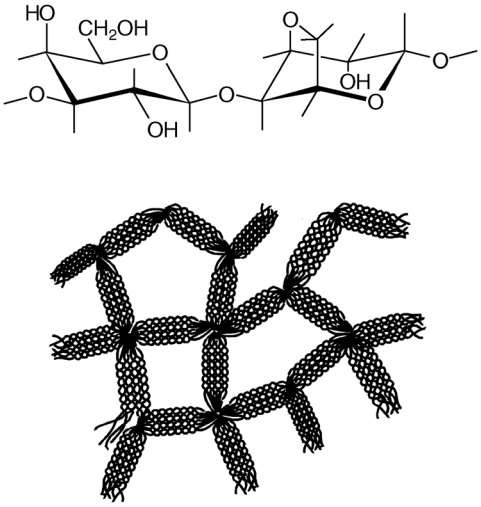


Figure 11.9 Structure of an agarose gel.

11.3.2 Gel Media for Electrophoresis

Advantages of Agarose Gels
They are nontoxic, simple to prepare, and ideal for the separation of high molecular proteins (>500 kDa). Because the pores are so large, such that immunoglobulins can diffuse into the matrix, specific detection of proteins via immune diffusion is enabled.

Disadvantages of Agarose Gels
They are never completely free of electroendosmosis, have a low retardation effect for proteins smaller than 100 kDa, and they are slightly opaque and fragile. Highly sensitive staining techniques like silver staining produce a high background.

Gel matrices can be prepared in the laboratory. However, ready-made gels and buffers in different sizes and for different methods are available from various suppliers. The preparation of porosity gradient gels is described in Section 11.3.5.

Agarose gels have relatively large pores: 150 nm pore diameter for a 1% (g ml^{-1}) to 500 nm for a 0.16% gel. Agarose is a polysaccharide that is produced from red seaweed by removal of the agaropectin. The agarose type is characterized by the melting temperature (35–95 °C) and the grade of electroendosmosis (m_r value), which is proportional to the leftover number of polar groups. Agarose is dissolved in boiling buffer and jells while it is chilling.

Thereby, the polysaccharide sol is converted into multiple helices, which form relatively thick fibers (Figure 11.9). This structure provides agarose gels with high mechanical stability and pores with large diameters.

The gels are usually prepared by pouring the hot agarose solution on a horizontal glass or plastic plate. The gel thickness results from the volume of the solution and the surface area. However, for some special applications it can also be poured into a pre-heated glass cassette to produce gels for vertical apparatuses.

Polyacrylamide gels are chemically inert and have a good mechanical stability. By copolymerizing acrylamide monomers with a crosslinker, mostly *N,N′*-methylenebisacrylamide (abbreviation Bis, Figure 11.10), a clear transparent gel is obtained with very low electroendosmosis. The pore size is defined via the total acrylamide concentration *T* and the

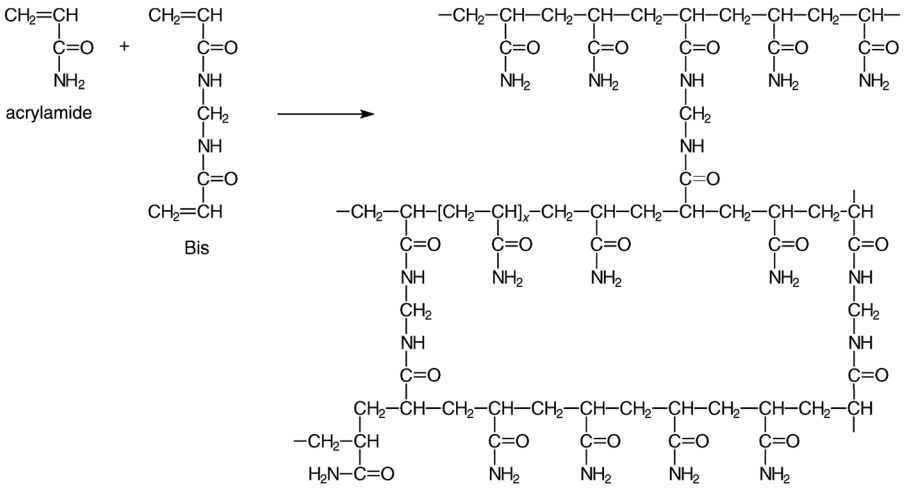


Figure 11.10 Structure of a polyacrylamide gel.

crosslinking factor C (g = gram weight):

$$T(\text{in }\%) = \frac{g_{\text{acrylamide}} + g_{\text{Bis}}}{100\text{ ml}} \times 100 \tag{11.12}$$

$$C(\text{in }\%) = \frac{g_{\text{Bis}}}{g_{\text{acrylamide}} + g_{\text{Bis}}} \times 100 \tag{11.13}$$

A gel with 5% T and 3% C has a pore diameter of 5.3 nm, a gel with 20% T has a pore diameter of 3.3 nm. The pore size decreases with increasing T at constant C. When the C value is increased or decreased at constant T the pore size will become larger; for 10% T gels the minimum lies at 5% C. However, gels with a too high crosslinking factor become brittle, opaque, and hydrophobic. They are only used in special cases.

The polymerization chain reaction is started by a radical donor, usually ammonium persulfate, which dissociates into radicals in presence of a tertiary amino group (*N,N,N′,N′*-tetramethylethylenediamine, TEMED). The reaction has to occur in the absence of air, because oxygen causes chain termination. The gels are prepared in vertical gel casters – gel rods in glass tubes, slab gels in cassettes – which are formed by two glass plates and spacers. Gels for horizontal systems are polymerized on a plastic film inside a cassette and removed from the cassette before use.

The polymerization effectiveness is dependent on temperature, pH value of the solution, acrylamide concentration, catalyst concentration, and on the concentration and kind of additives used. The gels should not be used shortly after preparation, because the matrix is only formed completely when the "silent polymerization" of the crosslinker has finished.

Advantages of Polyacrylamide Gels
They are mechanically stable and clearly transparent, have almost no electroendosmosis, provide good sieving effects over a wide separation range, and can easily be archived after separation and staining. A wide spectrum of staining methods can be applied.

Disadvantages of Polyacrylamide Gels
The monomers are toxic (skin and nerve toxin), the pore size is limited: proteins larger than 800 kDa cannot enter the gel. Basic gels have a short shelf-life due to alkaline hydrolysis.

11.3.3 Detection and Quantification of the Separated Proteins

The separated proteins zones can be directly stained in the gel. Usually, a fixation step must be performed before staining, either with 20% trichloroacetic acid for isoelectric focusing or a combination of 10% acetic acid and about 30% ethanol for electrophoresis gels. Agarose gels must be dried down on a glass or plastic plate after fixing prior to staining.

The most popular organic dye is Coomassie Brilliant blue, based on triphenylmethane, which was originally used for dyeing silk and wool. Several protocol variants are employed. The most sensitive detection in electrophoresis gels is obtained with colloidal Coomassie staining in the presence of ammonium sulfate and phosphoric acid. The superfluous dye is removed by washing the gel with water. For isoelectric focusing gels the most sensitive staining is obtained with acid violet 17. Depending on the dye-binding property of proteins, which can considerably vary, a sensitivity of detection in the region of 100 ng to 1 µg is obtained. Relative quantification of the protein zones can be performed over two orders of magnitude. Other dyes like Amido Black or Fast Green have lower sensitivities and they are therefore less frequently applied. Glycoproteins or lipoproteins can be detected with specific staining methods.

Silver staining is one of the most sensitive detection techniques used for proteins in gels. After fixing, several washing steps, and a crosslinking and sensitizing step the gels are submerged in silver nitrate or ammoniacal silver reagent some silver ions are reduced and – initiated by some functional groups and peptide bonds – bind to proteins and form silver nuclei. During the developing step the silver ions are reduced with formaldehyde to metallic silver. This reaction takes place considerably faster close to the silver nuclei than in the rest of the gel: the protein zones turn to dark brown and black. The reaction must be stopped at the optimum time point to prevent the background from turning black. Glycine solution affords the most efficient means of stopping, without further darkening. For isoelectric focusing gel the ammoniacal version is the method of choice. If protein zones have to be further analyzed with mass spectrometry, the crosslinking step with glutaraldehyde has to be omitted. The sensitivity of detection is down to less than 100 pg. The quality of the reagents as well as the water is highly critical. The disadvantages of the method are the high effort (many steps) and the very limited ability for quantification.

Western Blotting, Section 5.3.3

An interesting alternative for downstream analysis with Western Blotting or mass spectrometry is staining of the gel background instead of the proteins. This negative staining method is based on the generation of a salt complex formed by sodium dodecyl sulfate, imidazole and zinc sulfate in the gel, which results in a white opaque background and non-stained protein zones. The sensitivity of detection lies between that of Coomassie staining and silver staining. Unfortunately, quantification is impossible, because the proteins are not stained, only the gel background. The salt precipitate can be completely redissolved in a few minutes with EDTA solution.

Fluorescence staining with metal chelates like ruthenium(II) tris(bathophenanthroline disulfonate) or with complex organic molecules like epicocconone, now also available as a synthetic molecule, provide excellent quantitative results over a wide linear concentration range of more than four orders of magnitude. The sensitivity of detection comes close to that of silver staining. Special fluorescent dyes are available that specifically stain phosphorylated or glycosylated proteins.

It is also possible to prelabel the proteins prior to separation with fluorescent tags, which are either attached to primary amines or cysteine residues. Because no fixing and staining steps are necessary, prelabeling of proteins considerably simplifies and improves the detection workflow; after imaging the gels can be further processed with Western blotting, and it allows highly sensitive and quantitative visualization of the proteins. With some labeling approaches considerably higher sensitivity than with silver staining is obtained. When fluorescent tags are used that differ distinctly in their excitation as well as their emission wavelengths, differently labeled protein samples can be mixed together and separated together in one gel matrix. After fluorescence imaging at the respective different wavelengths the results of the different samples can be directly compared. This multiplexing method, known as DIGE (difference gel electrophoresis), is further described in Section 11.6.6.

To detect fluorescent dyes appropriate fluorescent scanners or camera systems are necessary. Most fluorescent stains and tags are compatible with subsequent mass spectrometry analysis.

Following native separation, very high sensitivity can be achieved with visualization methods that utilize specific properties of proteins: zymograms. Enzymatic activities can be detected directly by immersing the gel into a solution containing a specific substrate and coupling the reaction with a diazonium dye reaction.

If antibodies against the proteins of interest are available they can be used for specific detection with an immunological technique, such as immunofixation. This method works particularly well in agarose gels, and also in large pore size thin-layer gels. When narrow pore size gels are used, the proteins have to be transferred onto an immobilizing membrane before immunodetection. The advantages of immunodetection include improved sensitivity because of an amplification of the signal and the specific detection of the antigen in a mixture without prior purification.

By far the most sensitive detection methods are autoradiography and fluorography of radioactive labeled proteins. The protein bands are detected after exposure on an X-ray film or on a phosphor imager screen.

Imaging For documentation, qualitative and quantitative analysis, as well as visualization of fluorescent labeled or stained proteins either specific camera systems or densitometers are employed.

Lambert-Beer Law, Section 7.1.3

At present densitometers (also called scanners) are still technically superior to camera systems, particularly when larger gels have to be imaged. In principle a scanner uses a movable light source, and the light absorption is measured in every point of the gel. For one-dimensional gels a curve along the gel length following the extinction $\ln(I_0/I)$ is obtained, where I_0 is the irradiated light intensity and I the intensity measured in the detector. For two-dimensional gels the extinction is presented as a function of the gel area, like a topographic map: every position in the gel area is related to the extinction as a third dimension. For the analysis of visibly stained proteins scanners are used in the transparency mode. The reflection mode would not allow a reliable quantification of protein zones. The resolution of a densitometer depends on the width and the focusing of the light beam, and the increment of the measurements. Before measurement the instruments must be calibrated with the help of a photographic gray wedge.

The unit O.D. (optical density) is often applied in a biological and biochemical context and is specified as follows: 1 O.D. is the amount of substance, dissolved in 1 mL and measured in a 10 mm cuvette, giving an extinction value of 1. According to the Lambert–Beer law the extinction of a diluted solution increases linearly with its concentration. However, this correlation is no longer valid for bands and spots, because their protein concentrations are very high and staining intensities approach saturation. Above certain absorption values (O.D. >2.5 for white light and O.D. >4 for laser light) hyperbolic or sigmoidal functions can be observed.

Only relative quantification is possible, by comparing either increasing and decreasing signals or the intensity of a protein band in a sample with the intensity of a band of the same protein containing a known amount. In two-dimensional electrophoresis spot volumes are compared. Because the band widths inside a sample lane can sometimes vary, there are approaches in some software packages that treat one-dimensional lanes like two-dimensional separations: in this case the three-dimensional band volumes are compared.

For fluorescent detection laser sources with different wavelengths or white light sources, specific filters are used to adjust the light to the optimal excitation wavelength.

CCD camera documentation systems allow a faster imaging than scanners. They typically contain different light sources: white light for visible stains, UV light for ethidium bromide and other DNA fluorescent dyes like DNA stain and SYBR green, and LED lights for specific wavelengths for exciting other fluorophores. To reduce the background noise, CCD cameras are usually cooled with a Peltier element or liquid cooling.

The image analysis of two-dimensional gels is only possible with special software packages, which provide more or less automatic background subtraction, spot detection, spot matching or image warping, pattern comparison, quantification, statistical evaluation, and documentation. However, the quality of the results is still dependent on the quality of the separation pattern. Software cannot fix inadequate separation results and poor reproducibility.

11.3.4 Zone Electrophoresis

The separation principle of zone electrophoresis is based on the different migration velocities of charged particles in the electric field. The sample components' electrophoretic mobilities are dependent on their charge, size, and structure. In a highly restrictive separation matrix the influence of the particle size is higher than in a large pore medium. For the separation of nucleic acids, and in most cases for proteins, basic buffers are employed. The negatively charged particles migrate towards the anode. The *relative* electrophoretic mobility is determined by the co-migration of an ionic dye, like bromophenol blue or xylene cyanol, as a standard.

Owing to the different three-dimensional structures of proteins native electrophoresis does not allow the correlation of relative mobilities and molecular weights.

In practice field strengths of 10–100 V cm^{-1} are applied. The separation times are typically between 30 min and overnight, depending on the separation task and the design of the equipment. The relative electrophoretic mobilities are usually used as internal intermediate values. Typical values are rarely found in the literature, because too many parameters are necessary for the definition. Therefore, proteins are mostly defined by their size, molecular weight, or the isoelectric point.

The Ferguson Plot During electrophoresis in a restrictive gel the electrophoretic mobilities of proteins are dependent on, at the same time, both the amount of net charges and the molecule radius. Nevertheless it is possible to determine physicochemical parameters of proteins. The samples are separated under identical buffer, time, and temperature conditions, but in a series of gels with different concentrations. The resulting migration distances in these gels are different. After determining the relative mobilities, the m_r values of a protein are plotted logarithmically over the respective gel concentrations, which results in a straight line.

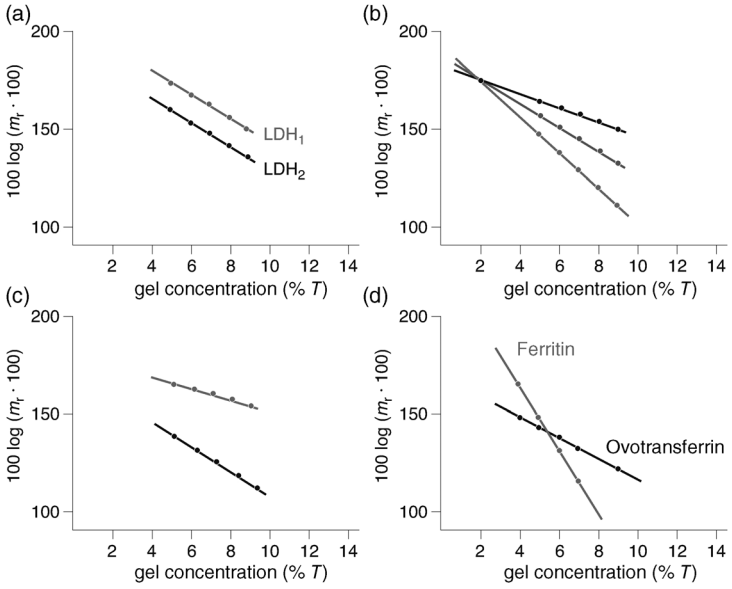


Figure 11.11 For the Ferguson plot the same samples are separated in native gels with different pore sizes. When the logarithms of the separation distances are plotted over the gel concentrations, straight lines result. For the interpretation of (a)–(d) see the text.

The slope of the line (Figure 11.11) is a measure for the molecule size and is defined as the retardation coefficient, the K_R value. For globular proteins there is a linear correlation between K_R and the molecule radius r (Stokes radius). This allows the molecule size to be calculated from the slope of the line. If the free mobility (in the absence of any restriction) and the molecule radius are known, it is possible to calculate the net charge.

For protein mixtures the following conclusions can be drawn from the position of the protein lines:

- Parallel lines indicate proteins of the same sizes, but different net charges, such as, for instance, in isoenzymes (Figure 11.11a).
- If several lines have a common intersection at a point located in the area $T < 2\%$, this means there are different polymers of a protein with the same net charges, but different sizes (Figure 11.11b).
- If there is no intersection between lines with different slopes, the protein of the upper line is smaller and more highly charged than the other one (Figure 11.11c).
- If the lines intersect in the area above $T = 2\%$, the protein, which intersects the y-axis further up, is larger and more strongly charged than the other one (Figure 11.11d).

11.3.5 Porosity Gradient Gels

The molecular size of proteins can also be determined with porosity gradient gels. Such gradients are obtained by continuously modifying the monomer concentration in the polymerization solution during gel casting.

If the monomer concentration and the crosslinking factor are high enough, the proteins will get stuck in the continuously closer meshed gel network at positions depending on their sizes. The migration velocities of the different protein molecules are dependent on their charges; therefore, the electrophoresis must be continued until the protein with the lowest net charge has reached its endpoint.

Structural proteins possess a larger volume than the tighter packed globular proteins of the same mass and get stuck at a larger pore size. Therefore, the migration distances cannot be correlated directly with the molecular weights. When a protein mixture is run perpendicular to the porosity gradient, similar information to that from a Ferguson plot can be obtained.

Several methods have been developed to prepare gels with linear or exponential pore gradients. The simplest technique is described here. Two solutions with different monomer concentrations are prepared. A gradient maker is used that contains two vessels that are connected with a channel at the bottom, which can be manually opened or closed. The system works according to the principle of communicating vessels: the height of both solutions stays

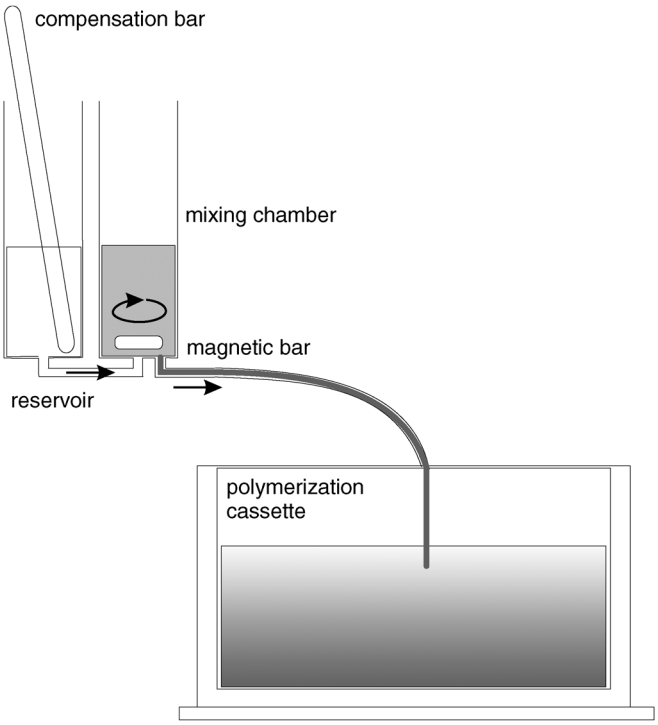


Figure 11.12 Casting of a linear gradient gel with a gradient maker. To stabilize the gradient the solution in the mixing chamber is weighted with sucrose or glycerol. There is a connecting channel between the two chambers. Half as much of the light solution flows from the reservoir into the mixing chamber as solution flowing out of the gradient maker into the polymerization cassette. In this way the solution has a continuously lower concentration; there is a continuous overlay of solutions.

balanced, therefore half as much of the light solution flows from the reservoir into the mixing chamber as solution flowing out of the gradient maker. With a plastic rod in the reservoir the volume of the magnetic bar and the difference in density is compensated. The low concentrated solution is continuously mixed into the higher concentrated solution, thus causing an increase in concentration from top to the bottom (Figure 11.12).

To avoid mixing of the layers glycerol or sucrose is added to the more concentrated solution. If an open mixing chamber is used, a linear gradient is obtained. For exponential gradients the mixing chamber is sealed with a stamp. The volume of the mixing chamber will remain constant during pouring of the solution: the same volume of light solution will flow from the reservoir to the mixing chamber as flows into the cassette.

11.3.6 Buffer Systems

For the analysis of proteins with isoelectric points in the acidic and neutral pH region homogeneous buffers like Tris-HCl or Tris-glycine pH 9.1, Tris-barbiturate and Tris-tricine pH 7.6 are used. For basic proteins acidic buffers like glycine–acetate pH 3.1 or aluminum-lactate pH 3.1 are employed; then the separation is directed towards the cathode. Owing to high net charges in continuous buffers the proteins have high mobilities, which can cause aggregation and precipitation of a part of the proteins during sample entry into small pore size polyacrylamide gels.

11.3.7 Disc Electrophoresis

Discontinuous electrophoresis solves several issues at the same time, such as protein aggregation during sample entry, and it produces sharper zones. The gel matrix is divided into two areas: the small pore size resolving gel and the large pore size stacking gel. Furthermore two different buffers are combined – mostly the Tris-chloride/Tris-glycine system according to Ornstein and Davis is applied. Therefore, this is described here as an example (Figure 11.13). The electrode buffer contains Tris-glycine, the resolving gel 0.375 mol l^{-1} Tris-HCl pH 8.8, and the stacking gel 0.125 mol l^{-1} Tris-HCl pH 6.8. This pH value lies in the vicinity of the isoelectric point of the glycine. Therefore, glycine is almost uncharged and has a very low electrophoretic mobility ("trailing ion"). The chloride ions in the gel buffers are highly charged and have a very high mobility ("leading ion"). The protein mixture is applied at the interface

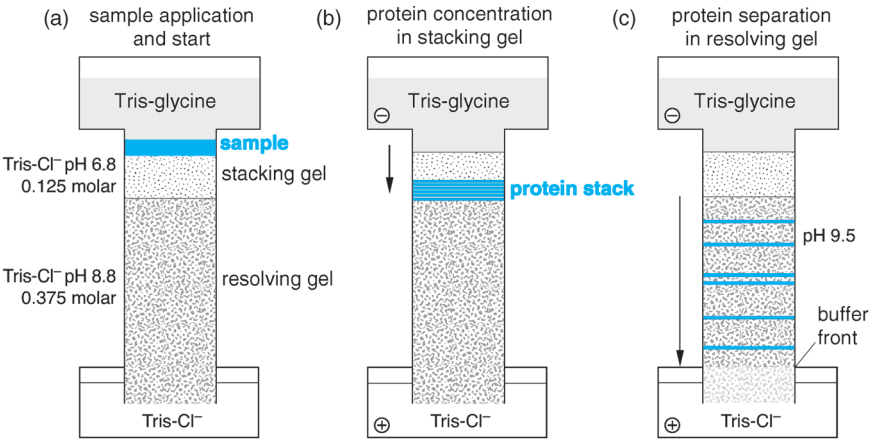


Figure 11.13 Principle of discontinuous electrophoresis: (a) The sample is applied on a large pore size stacking gel between chloride ions with high mobility and glycine ions with low mobility at pH 6.8. (b) In the electric field a stack of sample ions is formed during the migration of the ions, which will abruptly disintegrate after arriving at the boundary of the narrow pore size resolving gel. (c) Now the glycine ions are faster than the proteins, and separation occurs according to the conditions of zone electrophoresis, producing sharp protein zones.

between these ions: then the mobilities of the proteins lie in between those of the leading and the trailing ions. When the electric field is applied, all anions in such a discontinuous system start to migrate with the same velocity. This process is called isotachophoresis (which means in Greek "migration with the same speed").

None of the ions can migrate faster or slower than the others, as this would cause ion gaps between the ion zones and would interrupt the electric field. As a consequence a low electric field strength will automatically prevail in the area of ions with high mobilities, and a high field strength in the area of ions with low mobilities. Hence the protein ions are migrating inside a field strength gradient, and will form a stack of zones in the decreasing sequence of their mobilities ("stacking effect"). The ions with the highest mobility follow directly the leading ion (Cl^-), whereas those with the lowest mobility will migrate directly in front of the trailing ion (glycine). While migrating, the sample components are separated from each other and form a stack. Hereby the field strength gradient is converted into the shape of a stair. This process includes a regulation function: if a component migrates too rapidly into a zone of higher mobility it arrives in an area of lower field strength and will drop behind in this zone; in contrast, if a component migrates too slowly, it will be accelerated forward by the higher field strength. There is also a zone sharpening effect: a high concentration of the leading ion causes higher concentrations of the following sample components zones. This stacking effect has several benefits: it causes a slow migration of the proteins into the gel, preventing aggregation of the proteins, and there is a pre-separation and concentration of the zones.

This protein stack moves slowly with constant velocity in the direction of the anode until it arrives at the boundary to the resolving gel, which has a narrow pore size. Suddenly, the proteins experience a much higher friction effect, which considerably reduces their mobilities. However, the mobility of the small glycine molecules is not affected by the changing pore size. The glycine molecules overtake the proteins and experience a pH of 8.8 in the resolving gel. This causes a switch to high negative charges and hence an additional increase in mobility. The protein zones are now located inside a continuous buffer environment, and the conditions of zone electrophoresis apply: the stacks decompose and the component zones migrate with different velocities according to their electrophoretic mobilities. In addition, the sequence of protein ions will be rearranged, because in the narrow pore gel the molecule size has a markedly higher influence on the mobility, and the pH value rises to 9.5, due to the pK of the amino group of the buffering glycine.

In practice the stacking gel is polymerized onto the resolving gel shortly before electrophoresis to avoid mixing of the buffer ions due to diffusion.

In practice a mistake is frequently made: Because some widely used – misleading – protocols or operation procedures claim that the pH of the Tris-glycine running buffer should have a pH of 8.4, the running buffer is titrated down to this value with hydrochloric acid. This results at first in the exclusive migration of the highly mobile chloride ions, until the cathode reservoir is empty of chloride, and a disturbance of the stacking effect. As a consequence the running time can be very long (up to overnight), and the zones are poorly resolved. Therefore, measurement of the running buffer pH value should be abolished and titration of the running buffer must be abandoned.

Alternatively, disc electrophoresis buffer systems are applied as well, like, for instance, Tris-acetate/Tris-tricine or acetic acid-KOH/methoxyacetic acid-KOH/acetic acid-glycine, for cathodal separations, which are mentioned in the next section.

11.3.8 Acidic Native Electrophoresis

For the separation of basic proteins with pIs >6.8 an alternative buffer system must be used, because the proteins would migrate into the cathodal direction and be lost. There are proteins that can only be separated as positively charged ions in an acidic buffer system in the cathodal direction. One example is the separation of basic membrane proteins like prolamins for the cultivar differentiation of cereals. Figure 11.14 shows the band pattern of an acidic native gel electrophoresis of positively charged, alcohol soluble proteins (wheat gliadins) in a horizontal, urea-containing polyacrylamide gel. In this case a HEPES buffer pH 5.5 was used in the gel.

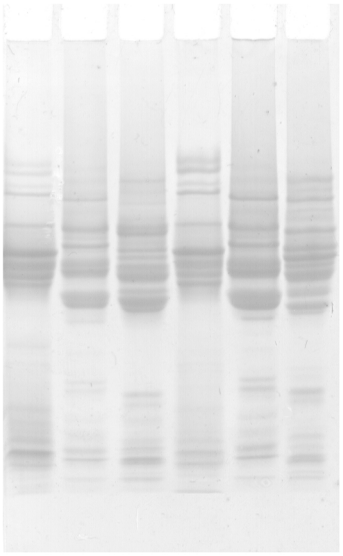


Figure 11.14 Acidic native polyacrylamide gel electrophoresis of alcohol protein extracts from different wheat cultivars. The gel was stained with a Coomassie Brilliant blue solution in acetic acid in the absence of alcohols.

11.3.9 SDS Polyacrylamide Gel Electrophoresis

In this method SDS (sodium dodecyl sulfate), an anionic detergent, is added to proteins prior to electrophoresis. SDS covers the intrinsic charges of polypeptide chains very effectively, resulting in micelles with constant negative charges per mass unit at a concentration of about 1.4 g SDS per g protein. Usually, the samples are heated to 95 °C with an excess of SDS for a few minutes, which causes disintegration of tertiary and secondary structures by cleavage of hydrogen bonds and stretching of the polypeptide chains. However, a complete stretching of the molecules can only be achieved by addition of thiol reagents like β-mercaptoethanol or dithiothreitol, which split the disulfide bonds between cysteines. This treatment will, though, cause the quaternary structures to fall apart into subunits. The amino acid chains loaded with SDS show a necklace structure and can be separated in an electrophoresis gel exclusively according to their *sizes*. This is a more correct definition than saying SDS electrophoresis separates according to molecular *weights*. An anionic tracking dye, for example, bromophenol blue, is added to the samples to determine the relative mobilities of the SDS-polypeptide bands.

SDS electrophoresis can be carried out in different basic buffer systems. However, the most frequently applied method employs the discontinuous Tris-HCl/Tris-glycine (Section 11.3.7) according to U.K. Laemmli, because it produces the best resolution: here 0.1% SDS is added to the running buffer.

As mentioned above, it is very important that the running buffer is not titrated with an acid. The gel can be polymerized without SDS, but the cathodal running buffer must contain sufficient SDS to replace the SDS molecules coming off the proteins during the run.

After SDS electrophoresis the logarithms of the molecular sizes of the polypeptides can be plotted against their relative mobilities. A sigmoidal shaped curve is obtained, which contains a linear section (Figure 11.15). The molecular sizes of the proteins or their subunits can be

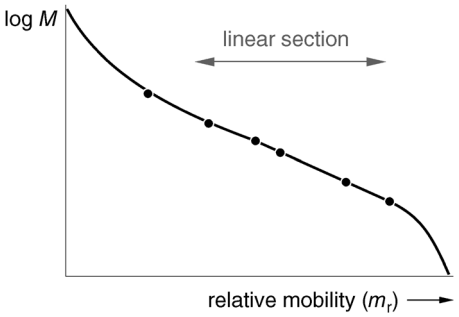


Figure 11.15 Molecular size curve after SDS polyacrylamide gel electrophoresis. The position of the linear section is dependent on the gel concentration.

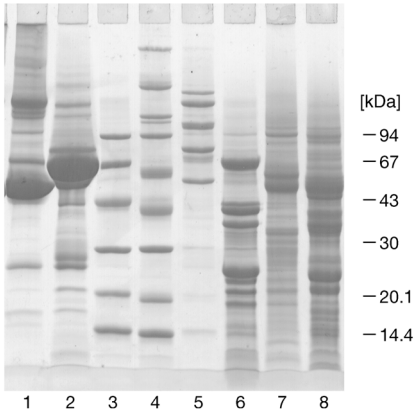


Figure 11.16 SDS polyacrylamide gel electrophoresis of human sera (tracks 1 and 2), plant seed extracts (6–8), and size standards (3–5). Staining with Coomassie Brilliant blue.

Advantages of SDS Electrophoresis SDS also solubilizes very hydrophobic and denatured proteins. Protein aggregations are prevented, because the surfaces of all protein–SDS micelles are negatively charged. The running times are shorter than for native proteins, because the charge density of SDS–protein micelles is higher. All sample components migrate in the same direction. The separation occurs according to only *one* parameter, the molecular size. Only one band per enzyme is obtained, because charge heterogeneities are cancelled.

Disadvantages of SDS Electrophoresis SDS denatures the proteins, most of them irreversibly. It cannot be applied for taxonomy studies, because protein isoforms caused by amino acid exchanges do not alter the molecular sizes. SDS is not compatible with several nonionic detergents that are used for the solubilization of some hydrophobic membrane proteins.

assessed by interpolation between the mobilities of protein standards with known sizes, which have been separated in the same gel. In the Tris-glycine buffer system proteins and peptides with molecular weights below 10 kDa are not resolved, because they co-migrate with the SDS bulk front.

For gels with constant T values the linear section of the curve stretches only over a limited section, which is determined by the proportion of molecular size versus gel pore diameter. For large pore size gels the section is shifted towards higher molecular sizes. For pore gradient gels the total separation range as well as the linear section is considerably wider. The bands are sharper, because a gradient gel counteracts diffusion.

For some studies, such as the analysis of serum and urinary proteins, the sample is not reduced to prevent the cleavage of immunoglobulins into subunits. However, in this case the polypeptides are incompletely stretched and migrate faster because of a decreased retardation in the gel; albumin with a size of 68 kDa runs with an apparent size of 54 kDa. Thus, it is only possible to determine molecular sizes of reduced subunits. It is, of course, possible to separate the same sample in one track in the reduced form and in another track in the non-reduced form to detect proteins with quaternary structures. Tracks 1and 2 in Figure 11.16 show, respectively, the band pattern of a non-reduced and a reduced serum sample.

SDS Electrophoresis for Low Molecular Weight Peptides In conventional buffer systems, like Tris-HCl/Tris- glycine, the resolution of peptides smaller than 10 kDa is very poor, because they are not destacked and co-migrate with the front. By using the method according to Schägger and von Jagow, where the trailing ion glycine is replaced by tricine, the gel buffer pH is changed to 8.4, the Tris concentration in the gel is increased up to 2½-fold, and a linear resolution in the range from 1 to 100 kDa is obtained.

Glycoproteins Glycoproteins bind much less SDS and therefore migrate more slowly than non-glycosylated proteins of the same size. When Tris-borate-EDTA buffer is used for sample preparation, the sugar moieties additionally become negatively charged, which accelerates the migration velocity.

11.3.10 Cationic Detergent Electrophoresis

Acidic electrophoresis in the presence of a cationic detergent is basically an inversion of SDS electrophoresis, and it denaturizes proteins less strongly. Thereby, the proteins are solubilized with cetyl(trimethyl)ammonium bromide (CTAB) or with benzyl(dimethyl)(*n*-hexadecyl)ammonium chloride (16-BAC) and applied on an acidic polyacrylamide gel, which also contains the respective detergent. Most often, a sodium or potassium buffer is used (see also the end of Section 11.3.7). The separation pattern is different from SDS electrophoresis, because the conformations of the polypeptides are different. The method is particularly suited for the analysis of membrane glycoproteins and is often employed as a first dimension combined with SDS electrophoresis for two-dimensional separations of very hydrophobic proteins.

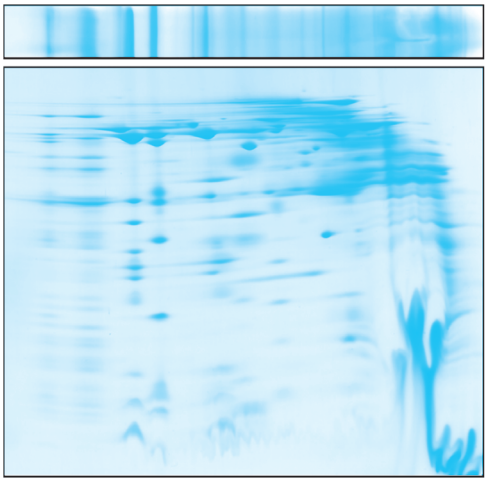


Figure 11.17 Two-dimensional Blue-native/SDS polyacrylamide gel electrophoresis. Separation of protein complexes under blue-native conditions was from left to right, separation of subunits of protein complexes in the presence of SDS was from top to bottom.

11.3.11 Blue Native Polyacrylamide Gel Electrophoresis

Because the anionic dye Coomassie Brilliant blue binds strongly to hydrophobic areas of proteins or protein complexes it can be used as a replacement for detergent during electrophoresis of intact protein complexes and membrane proteins. The complexes are solubilized with mild detergents like dodecyl maltoside and digitonin, mixed with Coomassie Blue G-250, and applied on a native pore gradient gel with a pH of 7.4. The separation is carried out at 4 °C. During the run the detergent on the complex surface is continuously replaced by the anionic dye. Analogously to SDS electrophoresis all the dye–protein complexes are negatively charged, and while they migrate towards the anode they become separated according to their sizes up to 10 MDa. However, in contrast to SDS electrophoresis the complexes are not modified or denatured; even super complexes remain intact. For further analysis the separation lanes are cut out; the gel strips are re-equilibrated in SDS buffer, and applied on a SDS electrophoresis gel. With SDS the complexes dissolve into subunits (Figure 11.17). The two-dimensional image delivers information on the composition of the respective complexes and super complexes.

11.3.12 Isoelectric Focusing

In isoelectric focusing (IEF) a protein or a peptide migrates electrophoretically through a pH gradient until it arrives at the pH value where its net charge is zero and, hence, its migration velocity is also zero. This is its isoelectric point. The net charge of a protein is the sum of all negative and positive charges on the amino acid side groups, whereby the three-dimensional structure of the protein also plays a role. In addition, phosphorylation, glycosylation, and oxidation status influence the charge condition. Some microheterogeneities in IEF patterns are related to such modifications of the molecules. The net charges of a protein plotted over a pH scale build a characteristic curve, which intersects the *x*-axis at the isoelectric point. When a protein mixture is applied on a point of the pH gradient, the different proteins exhibit different net charges, which cause them to migrate in the electric field in the direction of the opposite sign until they reach their isoelectric points. Because they no longer have a charge, they cannot migrate further. In contrast to other electrophoresis techniques, where the migration distances are influenced by the running time, IEF is an endpoint method. Furthermore, it implies a concentration effect, which acts against diffusion: if a protein were to diffuse away from its isoelectric point, it would become immediately charged and the electric field would move it back to the isoelectric point (Figure 11.18).

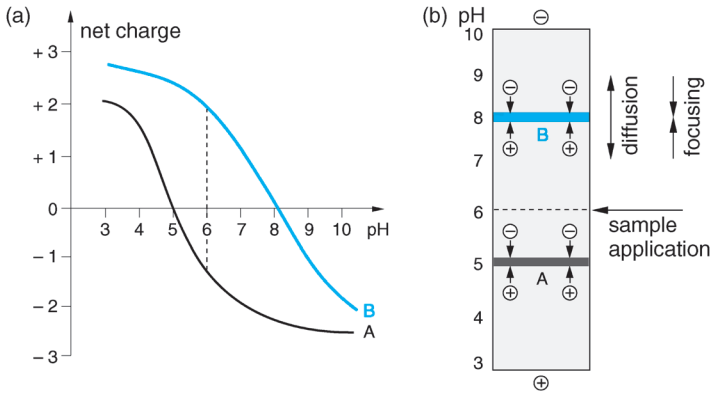


Figure 11.18 Principle of isoelectric focusing. (a) Net charge curves of two different proteins A and B. (b) When these two proteins are applied at a certain point of the pH gradient, they are either positively or negatively charged. In the electric field the proteins migrate towards their isoelectric point, where they cease migrating. The electric field focuses them on the isoelectric point, because in the vicinity of the pI they would gain a charge and, thereby, migrate back to it.

The resolution power of isoelectric focusing is defined as:

$$\Delta pI = \sqrt{\frac{D\left[\frac{d(pH)}{dx}\right]}{E\left[\frac{du}{d(pH)}\right]}} \qquad (11.14)$$

where: ΔpI is the resolution power,
D is the diffusion coefficient of the protein,
E is the field strength (V cm^{-1}),
d(pH)/dx is the pH gradient,
du/d(pH) is the mobility incline of the protein at the isoelectric point (see also titration curve analysis).

The ΔpI is the minimum pH difference necessary to resolve two neighboring bands. The equation allows conclusions about how to increase the resolution: narrow pH gradient intervals can be used for the separation of proteins with proximate isoelectric points. It also shows a limitation of isoelectric focusing in that the field strength can be increased by higher voltage, but is not unconstrained. The mobility incline of the protein at the isoelectric point cannot be influenced.

IEF can be performed on an analytical as well as a preparative scale. Preparative applications are described in Section 11.4.3.

Advantages of Agarose Gels
The separations are quicker, and proteins larger than 800 kDa can be separated. Agarose is nontoxic and does not contain any disturbing catalysts.

Advantages of Polyacrylamide Gels
Clear gels, less background staining, and less electroendosmosis, particularly in the basic area. In addition, denaturing gels containing urea can be prepared.

Separation Media Analytical IEF can be performed in polyacrylamide or in agarose gels. The best results are obtained in large pore size and very thin layer gels attached to a plastic film backing. Isoelectric focusing in agarose gels has only become possible since the intrinsic negative charges of agarose remaining from agaropectin have been counterbalanced by inbuilt positive charges. Usually 0.8–1% (w/v) agarose with very low electroendosmosis is used. Further properties of the two gel types are described in Section 11.3.2.

Measuring the pH Gradient The pH values have to be determined at the temperature of the separation, because they are highly temperature-dependent. Measurement with electrodes is quite challenging, because they have slow reaction times at low temperatures. Furthermore, carbon dioxide from the air diffuses into the gel and forms carbonate ions with the water. This results in decreased pH values, particularly in the range above pH 7. The interpolation of isoelectric points with the help of co-separated pI standard proteins and a calibration curve leads to fewer errors. The pI marker proteins are only applicable for native conditions, they cannot be used for pI determinations in urea gels, because the gradients are shifted, the proteins are dissociated into subunits, and their conformations are modified.

Kinds of pH Gradients The pH-gradient should stable and have a consistent conductivity and buffering capacity. These demands are fulfilled by two different concepts: (i) pH gradients built by free carrier ampholytes and (ii) immobilized pH gradients with fixed buffering groups in the matrix.

Carrier Ampholytes These are heterogeneous mixtures of synthesis products consisting of several hundred different homologues of small molecular aliphatic oligoamino-oligocarbonic acids, which differ in their isoelectric points. An ideal carrier ampholyte has the following properties:

- high buffering capacity and solubility at its pI,
- good and consistent conductivity at its pI,
- an absence of biological effects,
- low molecular mass.

The different amphoteric homologues should have the same concentrations. Gaps in the pH spectrum must not exist. Natural ampholytes like amino acids or peptides are not applicable, because they exhibit low buffering capacity at their isoelectric points.

Usually, gels with 2% (w/v) carrier ampholytes are employed. Initially, the gels have a uniform mean pH value. At this pH all carrier ampholytes are charged: the basic ones positively, the acidic ones negatively. When an electric field is applied, a pH gradient develops: the negatively charged migrate towards the anode, the positively charged towards the cathode. This causes the anodal side of the gel to become more acidic, and the basic side to become more basic. The carrier ampholyte molecules with the lowest pI migrate to the anodal end, the most basic ones to the cathodal end of the gel. All the others arrange themselves in between in the order of their pIs and render this pH value to their environment in the gel. In this way a stable, continuous pH gradient is created.

At their pIs the carrier ampholytes lose their net charges, and the conductivity in the gel decreases. As the pH gradients are temperature dependent, the separation must be performed at a defined temperature.

Advantages of Carrier Ampholyte IEF
There is a choice between agarose and polyacrylamide gels. The gels are easy to prepare. Carrier ampholytes act as zwitterionic buffers, they keep proteins in solution.

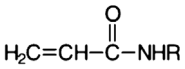
Disadvantages of Carrier Ampholyte IEF
The composition – and hence – the profile of the pH gradient is not completely reproducible due to the complex synthesis of the carrier ampholytes. The formation of adducts with some sample components has been reported, such as with heparin.

Electrode Solutions To keep the gradient stable, in most cases electrode strips soaked with electrode solutions are placed between the gel edges and the electrode wires – an acid at the anode and a base on the cathode. When an acidic carrier ampholyte molecule arrives at the anode, its basic group becomes positively charged and gets pushed back into the gel.

Separator IEF If the resolution between some proteins is not sufficient, separators can be added: amino acids or amphoteric buffer substances, which flatten the pH gradient in the vicinity of its isoelectric point. With this measure, closely neighboring protein bands can be completely separated; for example, glycosylated hemoglobin can be separated from the neighboring hemoglobin band by adding 0.33 mol l^{-1} β-alanine when IEF is carried out at 15 °C.

Cathodal Drift During a long focusing time the gradient starts to drift in both directions (also called plateau phenomenon), but mainly in the cathodal direction. This can cause losses of basic proteins.

Immobilized pH Gradients Immobilized pH gradients (IPGs), which are completely different, are grafted into the polyacrylamide matrix. This is achieved with so-called immobilines, which are not amphoteric but bifunctional and have the following chemical structure:



where R contains a buffering group, either a carboxylic or a tertiary amino group, as shown in Table 11.1.

These immobilines are acrylamide derivatives and are at the same time both weak acids or weak bases, which are defined by their p*K* values. To buffer a certain pH value, at least two different immobilines are needed: an acid and a base. Figure 11.19 shows a schematic structure of a polyacrylamide gel with copolymerized immobilines; the pH value results from the mixing ratio of the immobilines. pH gradients are obtained by continuously modifying the mixing ratio during casting the gel, analogously to pore gradient gels (Section 11.3.5). The physicochemical

Table 11.1 Chemical structures of acidic and basic acrylamide derivatives for the preparation of immobilized pH gradients.

p*K*	Chemical Structure
3.6	CH_2=CH–CO–NH–CH_2–COOH
4.6	CH_2=CH–CO–NH–$(CH_2)_3$–COOH
6.2	CH_2=CH–CO–NH–$(CH_2)_2$–N⟨O⟩
7.0	CH_2=CH–CO–NH–$(CH_2)_3$–N⟨O⟩
8.5	CH_2=CH–CO–NH–$(CH_2)_2$–$N(CH_3)_2$
9.3	CH_2=CH–CO–NH–$(CH_2)_3$–$N(CH_3)_2$
10.3	CH_2=CH–CO–NH–$(CH_2)_3$–$N(C_2H_5)_2$
> 12	CH_2=CH–CO–NH–$(CH_2)_2$–$N(C_2H_5)_2$

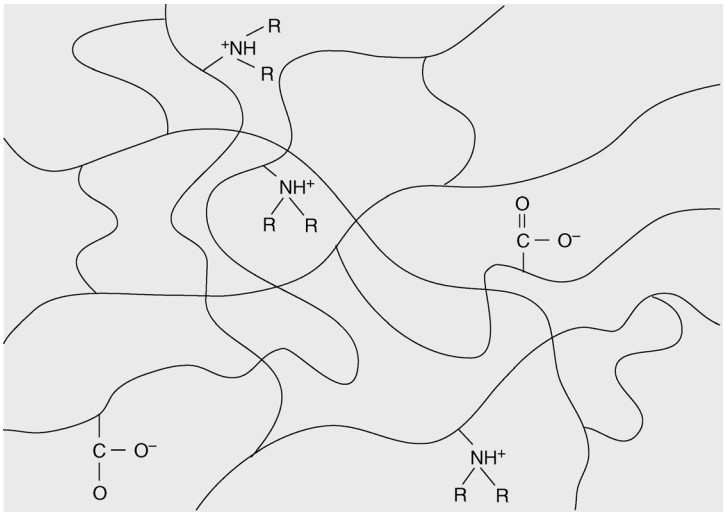


Figure 11.19 Immobilized buffering groups in a polyacrylamide gel network. With certain concentrations of the respective weak acids and bases defined by their p*K* values the pH values can be adjusted exactly.

principle is an acid–base titration, with the respective pH value defined by the Henderson-Hasselbalch equation:

$$\mathrm{pH} = \mathrm{p}K_B + \log\left(\frac{c_B - c_A}{c_A}\right) \tag{11.15}$$

when the buffering immobiline is a base; c_A and c_B are the molar concentration of the acidic and basic immobilines, respectively. If the buffering immobiline is an acid, the equation looks as follows:

$$\mathrm{pH} = \mathrm{p}K_A + \log\left(\frac{c_B}{c_A - c_B}\right) \tag{11.16}$$

Advantages of IPG-IEF
Immobilized pH gradients are absolutely time-stable, there is no cathodal drift or plateau phenomenon. It is possible to prepare exactly engineered pH gradients, with very broad as well as very narrow intervals. Extremely high resolutions of proteins up to pH 12 can achieved. Immobilized pH gradients are particularly suited for the first dimension of two-dimensional electrophoresis.

Disadvantages of IPG-IEF
The preparation of IPG gels is more complicated than carrier ampholyte gels. IPGs require high field strength and separation time. They are less suitable for native IEF.

Preparation of Immobilized pH Gradients In practice immobilized pH gradients are prepared by linear mixing of two different monomer solutions with a gradient maker (Figure 11.12). Both solutions contain acrylamide monomers, crosslinker, and catalysts for polymerization of the gel matrix. The heavier solution, containing glycerol, is adjusted to the acidic, the light solution is adjusted for the basic end of the gradient. During polymerization the immobiline molecules become covalently attached to the polyacrylamide network. The gels are polymerized onto a plastic film. Because the conductivity of fixed pH gradients is very low, the catalysts have to be removed after polymerization by washing the gels with distilled water. Afterwards the gels are dried. Before use the gels are rehydrated with an aqueous solution of all additives needed for IEF, like urea, dithiothreitol, carrier ampholytes, and detergents. Figure 11.20 shows the pherogram of

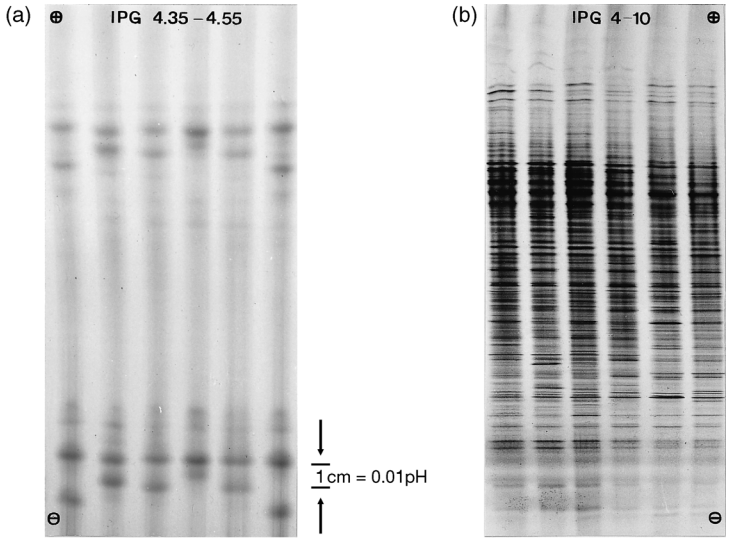


Figure 11.20 Isoelectric focusing in immobilized pH gradients. (a) Separation of α_1-antitrypsin PiM subtypes of human sera. Ultra-narrow IPG 4.35–4.55, 20 cm, 20% glycerol; ΔpI = 0.001. (b)Separation of seed proteins of different cultivars of broad beans (*Vicia faba*). Wide IPG 4–10, 20 cm, 6 mol l^{-1} urea, 15% glycerol. Source: Görg, A. *et al*. (1986) *Electrophoresis 86* (ed. M.J. Dunn), VCH, Weinheim, pp. 435–449. With permission, Copyright © 1986 Wiley-VCH Verlag.

an IEF in a very narrow pH gradient (IPG 4.35–4.55) for phenotyping of α_1-antitrypsin variants (a) and a wide pH gradient (IPG 4–10) for the taxonomy of bean seed cultivars (b).

With IPGs it is possible to pre-calculate and cast any pH gradients, which can be modeled according to the separation task. It is possible to obtain extremely high resolution by preparing very narrow pH intervals, down to 0.01 pH units per cm ($\Delta pI = 0.001$). In addition, very wide linear or nonlinear gradients from pH 2.5 to 12 can be prepared. Because the gradient is covalently bound to the gel, it remains unchanged during the entire separation period. The reproducibility of the protein patterns is very high.

Immobilines are defined single substances, not mixtures of compounds. The profile of an immobilized pH gradient is not influenced by proteins or salt in the sample, the iso-pH lines are perfectly straight. IPG gels are the only electrophoresis gels that can be cut into narrow strips and run without any edge effects. IPGs are mainly used for the denaturing first dimension of high resolution 2D electrophoresis as narrow film-backed strips. Ready-made IPG strips are provided by different companies.

Titration Curve Analysis With this simple method the net charge curves of proteins can be displayed. A quadratic carrier ampholyte gel with a narrow groove preformed in the gel surface is needed. First an IEF without sample is carried out in the length direction of the groove until the pH gradient is established. Then the gel is turned by 90°. The sample is applied into the groove. When an electric field is then applied perpendicular to the pH gradient, the focused and uncharged carrier ampholytes remain in their positions, but the proteins migrate according to their net charges – dependent on the respective pH value – and form their titration curves (Figure 11.21). In principle these are multiple parallel native zone electrophoreses in a single gel. The isoelectric point of a protein lies at the position where the titration curve intersects the groove. With this analysis interesting information on the properties of a protein is obtained, such as on the mobility incline at the isoelectric point and on conformation changes or binding of ligands dependent on the pH value. Furthermore, the pH optima for native electrophoresis and for protein elution in ion exchange chromatography can be assessed.

Advantages of IEF
The method exhibits a very high resolution power and is an endpoint method. Genetic differences are detected with high sensitivity: proteins differing by the exchange of one amino acid can be distinguished. The important physicochemical parameter, the isoelectric point, can be directly determined. IEF can readily be combined with other techniques, like SDS gel electrophoresis.

Disadvantages of IEF
Some proteins, for example, membrane proteins tend to aggregate and do not enter the gel, particularly under native conditions. Very basic proteins cannot be easily focused with carrier ampholyte IEF. Separations take longer time than in zone electrophoresis. IEF gels require an intensive fixation of the proteins with trichloroacetic acid, in order to elute the carrier ampholytes and to prevent elution of the proteins.

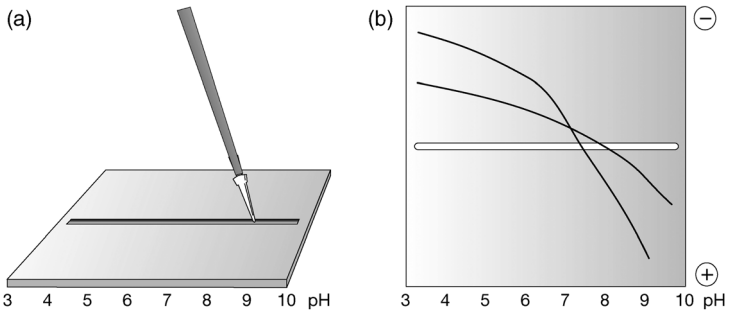


Figure 11.21 (a) For titration curve analysis the sample is applied on a prefocused carrier ampholyte gel, which contains a pH gradient. (b) In the electric field perpendicular to the pH gradient the proteins migrate according to their charges and form their own net charge curves.

11.4 Preparative Techniques

11.4.1 Electroelution from Gels

In many cases the amount of protein in a band is sufficient for further analysis. Elution of proteins from a high resolution polyacrylamide gel can only be performed with an electrophoretic method.

Figure 11.22a shows a simple principle – the cut out gel piece containing the protein fraction is placed on a frit in a glass tube. The end of the tube is closed with a dialysis membrane. The tube is inserted into a vertical electrophoresis chamber, which allows several electroelutions to be performed at the same time. The protein is electrophoretically transported towards the dialysis membrane, which does not allow a further passage. Most often, ammonium carbonate buffer is used, which evaporates into the gas phase during lyophilization. The advantage of this method is the use of standard equipment. Admittedly a dialysis membrane has to be used, which can irreversibly adsorb some problems.

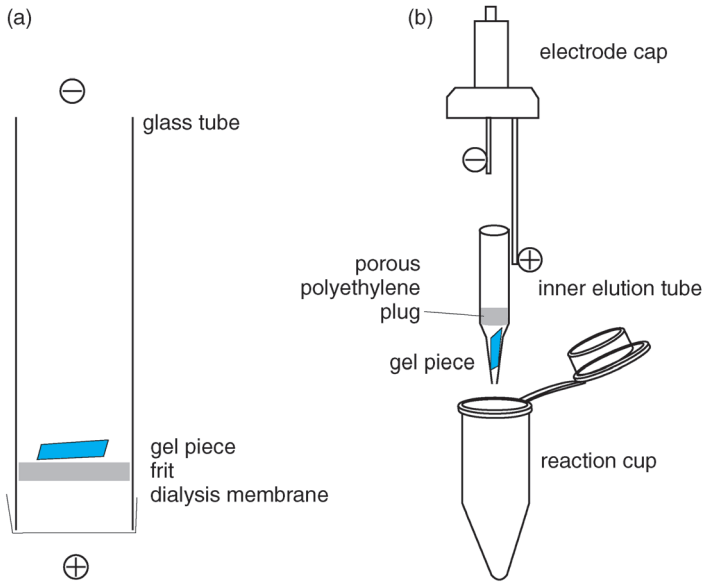


Figure 11.22 Electroelution. (a) The gel piece containing the protein fraction is placed on a frit in a glass tube, which is closed at the lower end with a dialysis membrane. For elution the glass tube is inserted into a vertical electrophoresis chamber. (b) Here the gel piece is placed in a conical inner elution tube, which is closed with an electrode cap. This system works without a dialysis membrane.

The method shown in Figure 11.22b works without any membrane. Here a special conical elution tube is inserted into a reaction cup (1.5 ml volume) with a cut-off tip. After placing the gel piece into the elution tube it is closed with a porous polyethylene plug. The buffer is filled in and the electrode cap is applied, which contains a cathodal electrode for the elution tube and an anode for the reaction cup. This mini-appliance is inserted into a specially designed electroelution chamber, in which several electroelutions can be carried out in parallel. The purified protein is extracted from the reaction cup. The advantage is that it works without a membrane. However, a special electrophoresis chamber is required.

11.4.2 Preparative Zone Electrophoresis

Electrophoretic techniques are in general characterized by their high resolution power. The amount of proteins to be separated is usually limited to a few milligrams, mainly because of inadequate dissipation of Joule heat. In principle an appliance for preparative zone electrophoresis consists of a glass tube, which contains a resolving gel. The zones arriving at the lower end are transported to a fraction collector via a continuous buffer stream (Figure 11.23). The different designs are distinguished by the type of cooling (e.g., jacket cooling) and the means of sample extraction.

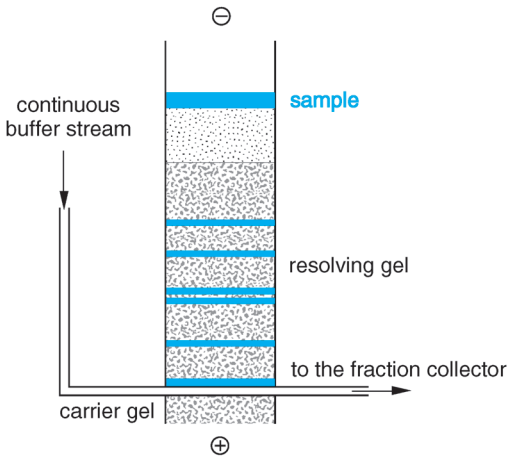


Figure 11.23 Principle of preparative electrophoresis: The protein zones arriving at the end of the vertical gel are transported to a fraction collector with a buffer stream.

11.4.3 Preparative Isoelectric Focusing

In isoelectric focusing high field strength can be achieved with low conductivity. Therefore, the cooling issues are less than in zone electrophoresis. In the following, three different techniques are described: IEF in granulated gels with carrier ampholytes, IEF in free solution between isoelectric membranes, and off-gel isoelectric focusing. A fourth technique, free flow electrophoresis, is presented in Section 11.6.

Preparative Carrier Ampholyte IEF Here a highly purified dextran gel is mixed with carrier ampholytes and poured into a horizontal trough. After prefocusing to establish the pH gradient, at a certain position a part of the gel is dug out with a spatula, mixed with the sample, and applied back at the same position. After IEF the protein or enzyme zones are detected with a paper print, which is stained with Coomassie Blue or a zymogram substrate reaction (like a gel). Then the zone containing the protein of interest is dug out (see also prefractionation Section 11.6.2). Alternatively, the entire gel is fractionated with a grid; the single fractions are eluted from the gel in small tubes containing nylon meshes. The carrier ampholytes are removed with gel filtration, ultrafiltration, dialysis, and ammonium sulfate precipitation of electrophoresis. In this way protein amounts of up to 100 mg can be purified.

There is also a matrix-free system for preparative fractionation of proteins based on their pI using liquid-phase IEF in a rotating tube divided into 20 compartments by woven polyester screens. The advantage is that the fractions are available in free solution, the disadvantage is that proteins precipitating at their pIs can block the dividing screens.

Preparative IEF between Isoelectric Membranes This technique is based on a continuation of the principle of immobilized pH gradients. Instead of a gel with an immobilized pH gradient a multichamber appliance is used, which is segmented by buffered isoelectric polyacrylamide-membranes (Figure 11.24). Isoelectric membranes can be easily prepared: glass fiber membranes are soaked in an acrylamide-immobilines polymerization solution with a defined pH value and polymerized under exclusion of air. The pH values needed are determined by an analytical pre-experiment in an IEF in an immobilized pH gradient. After the polymerization the pH value is grafted to the membrane. The pH values for the membranes are selected in such a way that the isoelectric point of the protein to be purified is enclosed as tightly as possible. For instance:

$$pI_{protein} = 6.15 \rightarrow pH_{membrane\ I} = 6.10; \quad pH_{membrane\ II} = 6.20$$

When a protein mixture is deposited in this compartment and an electric field is applied, the components with higher and lower isoelectric point migrate through the membranes into the neighboring compartments. The protein to be purified remains in the compartment between the two membranes and can be removed from there. With this method up to a gram of protein can be purified.

In off-gel isoelectric focusing a narrow gel strip containing an immobilized pH gradient (IPG strip; Section 11.6.3) is laid into a horizontal chamber, and a fractionating frame with 24 compartments is placed directly onto the gel surface. Then 150 µl of a diluted sample is pipetted

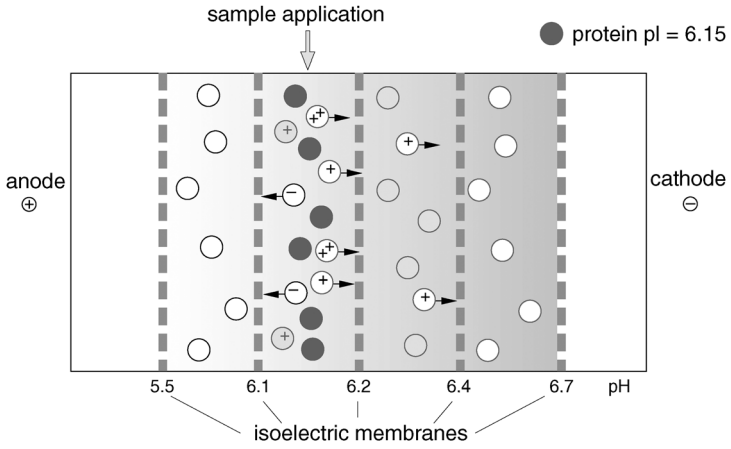


Figure 11.24 Principle of preparative IEF between isoelectric membranes. In the electric field the proteins with higher or lower isoelectric points migrate through the membranes into the neighboring compartments until they are located between membranes of enclosing pH values. They stay in the compartments where they are isoelectric and can be taken out from there at the end of the separation.

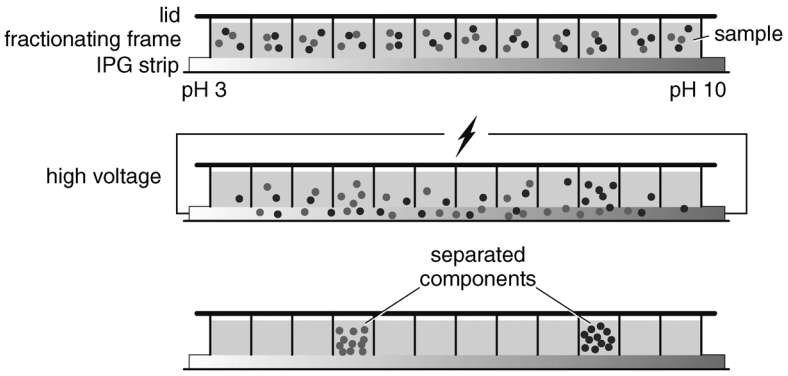


Figure 11.25 Principle of off-gel isoelectric focusing: The sample components are separated and enriched according to their isoelectric points in the liquid phase at the surface of an IPG strip. The diluted sample is pipetted into small compartments of a fractionating frame. In the electric field the charged proteins or peptides migrate through the gel layer of the IPG strip into the next chamber, until they reach the pH value corresponding to their pIs.

into each compartment. The gel surface of the IPG strip seals the bottom of each single compartment. The top of the small chambers are sealed with a lid (Figure 11.25).

When a high voltage is applied on the ends of the IPG strip, the charged proteins and peptides migrate through the gel layer until they reach the compartment corresponding to their pIs. The protein and peptide fractions are in aqueous solution and can be drawn out from the compartment. The major advantages of off-gel IEF are the relatively large protein amounts (up to a mg) that can be fractionated and the availability of the protein fractions in liquid solution.

11.5 Free Flow Electrophoresis

In free flow electrophoresis separation occurs during a continuous flow in a thin buffer layer between two parallel plates. The sample is applied at a defined position on one side of the flow chamber. An electric field is applied perpendicular to the laminar buffer flow in the flow chamber flanked by electrodes. In this way charged sample components are deflected perpendicular to the flow direction according to their different electrophoretic mobilities. The separated fractions arrive at defined positions at the end of the flow chamber and are collected by an array of thin tubing (Figure 11.26).

Free flow electrophoresis distinguishes itself from the electrophoresis techniques described above in two significant ways:

- Because there is no gel or other stabilizing matrix it is also possible to separate particles like cell organelles or entire cells, viruses, and bacteria.
- This is a continuous technique: buffer and sample flow through the apparatus perpendicular to the electric field and the separation direction.

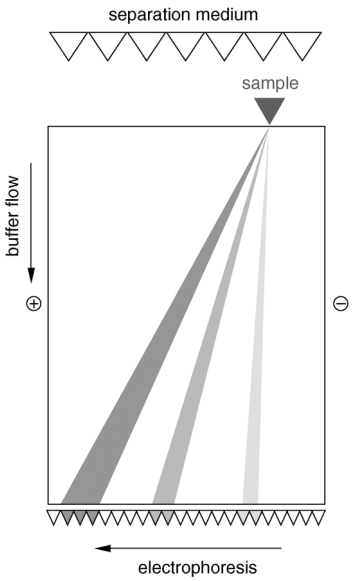


Figure 11.26 Principle of free flow electrophoresis. A continuous thin buffer layer flows through a separation chamber. The electric field perpendicular to the flow direction deflects the sample components differently according to their mobilities. The different fractions arrive at the end of the flow chamber at different, but at constant positions.

The flow cell is mostly set up in a vertical direction, the gap between the plates is very thin (from 1 mm to less than 200 µm). Modern systems allow the application of zone electrophoresis, isotachophoresis, and isoelectric focusing. Special effects can be achieved with field jump electrophoresis and modifications of isoelectric focusing with added amphoteric substances.

In field jump electrophoresis different field strength are adjusted by employing buffers with different ionic strength. The sample solution is introduced in a broad zone via the central inlets, the buffer solutions at the right- and left-hand side have a 20-fold higher conductivity than the sample solution. The sample ions will be strongly deflected towards the cathode or the anode depending on their charges. When they arrive at the boundary between sample and buffer flow, their electrophoretic mobility is considerably decreased, which leads to an enrichment of the sample ions at this interfaces.

Free flow isoelectric focusing uses either carrier ampholytes or multicomponent buffers composed of amphoteric and non-amphoteric substances. When a multicomponent buffer is used it is not possible to obtain a linear pH gradient, but consumable costs are considerably lowered. The gradients achieved by a mixture of carrier ampholytes or buffers are usually called

natural pH gradients, whereas artificial pH gradients are obtained by buffers flowing through the chamber in different pH steps in parallel zones.

11.6 High-Resolution Two-Dimensional Electrophoresis

When two electrophoretic methods according to orthogonal separation parameters are combined, like pIs in the first and molecular sizes in the second dimension, a high resolution can be achieved. However, only when both dimensions are carried out under fully denaturing conditions can it provide the highest resolution for the analysis of complex protein mixtures (e.g. proteomes of cells, tissue, and body fluids). It is possible to separate several thousand protein spots in one gel. These conditions are achieved by the presence of uncharged chaotropes, like urea and thiourea, reduction reagents, carrier ampholytes, and zwitterionic detergents during sample preparation and in the focusing gel. In the original technique the IEF step was performed in individual gel rods in thin glass tubes with carrier ampholyte generated pH gradients. This method requires a very high level of skill and has therefore been replaced in most laboratories by IEF in film-backed IPG strips (see immobilized pH gradients in Section 11.3.12). IPGs have not only simplified the method but also improved the reproducibility of the 2D patterns, because of the fixed pH gradients. The second dimension occurs in SDS polyacrylamide gels, either in a horizontal flatbed chamber or in a vertical tank.

For the complete separation of cell lysates, tissue extracts, and body fluids, gels in the size range 20×20 cm or larger are employed. The schematic workflow of high resolution two-dimensional electrophoresis according to Görg *et al.* is shown in Figure 11.27. In principle, representation of the complete protein content of the sample would be desirable. This can be done with a wide pH gradient (e.g., pH 3–11) (Figure 11.28). Most protein mixtures are so highly complex that some of the spots contain several different proteins. For the quantification of a protein it is important that one spot does not contain more than one protein. A complete resolution of the proteins can either be achieved by increasing the gel size or by partition of the pH scale into shorter pH intervals.

Mass Spectrometry, Chapter 15

Proteomics, Chapter 40

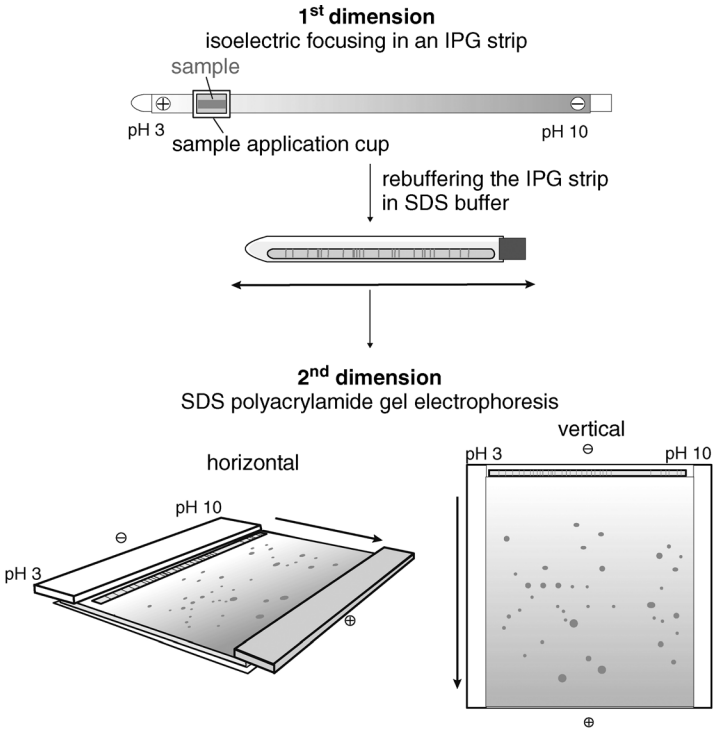


Figure 11.27 Workflow of high resolution 2D electrophoresis with an IPG strip. Horizontal IEF in a film-backed IPG strip. Equilibration of the IPG strip. Application of the IPG strip on a horizontal or a vertical SDS gel. Source: Görg, A. *et al.* (1988) *Electrophoresis*, 9, 531–546. With permission Copyright © 1988 Wiley-VCH Verlag.

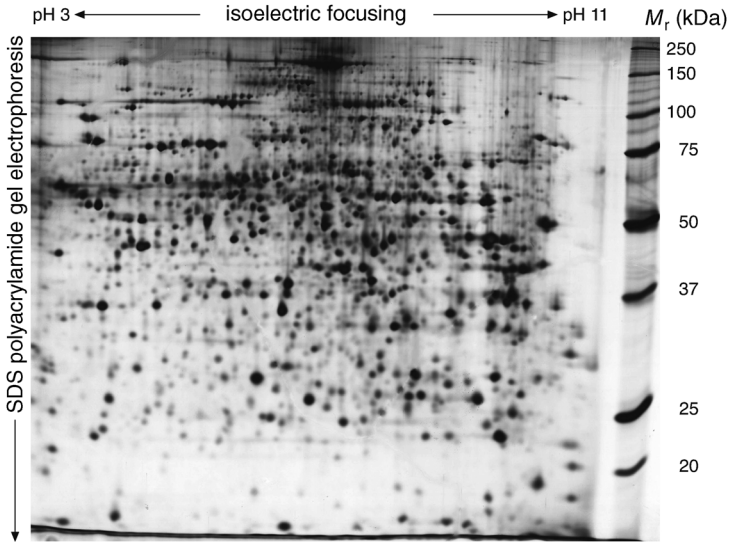


Figure 11.28 High resolution 2D electrophoresis of mouse liver proteins. First dimension: IPG 3–11, second dimension SDS polyacrylamide gel electrophoresis, silver staining. Source: Görg, A. (2000) *Electrophoresis,* 21, 1037–1053. With permission, Copyright © 2000 Wiley-VCH Gmbh & Co. KGaA.

Furthermore, a gel with a narrow pH gradient exhibits a higher loading capacity. This allows application of higher amounts of proteins to detect also molecules with low copy numbers. Figure 11.30 below shows the separation of mouse liver proteins in the narrow interval pH 6–7.

11.6.1 Sample Preparation

Two-dimensional electrophoresis is mainly used for the separation of highly complex protein mixtures, therefore appropriate sample acquisition and preparation are the most important prerequisites for correct results. During cell lysis and tissue extraction any modifications of the protein mixture caused by the activities of enzymes must be precluded. Immediate heat stabilization of tissue samples by controlled conductive heating permanently prevents biological changes to maintain the protein properties and composition. When working with cell lysates it is important to work rapidly, because the addition of protease inhibitors will not completely prevent spontaneous protein digestion. Ultracentrifugation prior to sample application removes solid cell debris, which can cause clogging of gel pores. Very often the sample contains several compounds that disturb analysis, like lipids, polysaccharides, nucleic acids, salt ions, and solid particles such as cell wall material. Furthermore, the formation of complexes has to be avoided, or the complexes have to be disintegrated. Precipitation with methanol and chloroform, or with a combination of trichloroacetic acid, deoxycholate, sodium hydroxide, and acetone, is a very powerful way to remove all disturbing substances at once and to dissolve complexes. Precipitation is the only measure that irreversibly inhibits protease activities. Resolubilization is performed in an aqueous solution of high molar urea and thiourea, zwitterionic detergent, thiol reagent, and carrier ampholytes. In general, nonionic detergents are no longer used, because they are not compatible with mass spectrometry; they have been almost completely replaced by zwitterionic detergents.

11.6.2 Prefractionation

There are various reasons for prefractionation; a few examples are given here.

Subcellular Components One and the same protein can have different functions in a cell, depending on which organelle it is located in. Topographic information can be attained when cell organelles are separated via density gradient centrifugation, detergent fractionation, or free flow electrophoresis. Furthermore, this step reduces the complexity of the protein mixture.

Affinity Chromatography This reduces the complexity of a protein mixture and removes highly abundant proteins from special samples, for example, human serum or plasma.

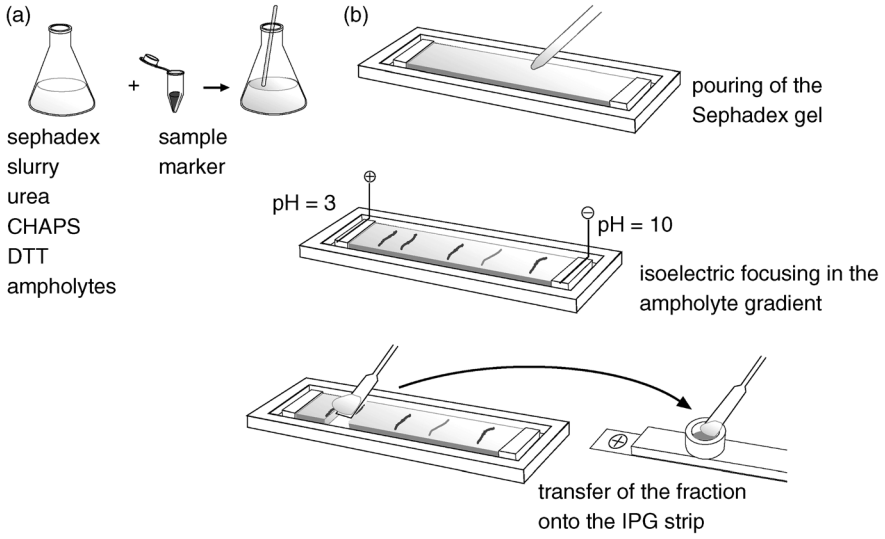


Figure 11.29 Schematic drawing of prefractionation of a protein mixture by IEF in a carrier ampholyte Sephadex (dextran) gel. (a) Preparation of the gel slurry with the protein mixture. (b) Low molecular weight amphoteric dye markers provide easy assessment of pH values for the correct withdrawal of the protein fractions. The gel fractions containing the focused proteins are applied directly on the IPG strip. Source: Görg, A. *et al.* (2002) *Proteomics*, 2, 1652–1657. With permission, Copyright © 2002 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

Fractionation according to Charge When the complete protein mixture is applied on an IPG strip with a narrow pH interval, the proteins with pIs outside this interval accumulate at the electrodes. These proteins are lost for further analysis, disturb isoelectric focusing, and limit the loading amount. Prefractionation according to the isoelectric points allows the further analysis of all proteins and increases the protein loading capacity. There are four different methods: free flow electrophoresis (Section 11.5), electrophoretic separation between isoelectric membranes (Section 11.4.3), liquid-phase carrier ampholytes IEF in a rotating tube, and isoelectric focusing in a horizontal granulated gel. The latter method can be performed in a short time with no extra equipment. The principle of this technique is shown in Figure 11.29. Highly purified dextran ("Sephadex") is soaked in the presence of all additives needed for denaturing IEF: urea, CHAPS, dithiothreitol (DTT), carrier ampholytes, and colored low molecular weight amphoteric pI standards. The protein solution is added immediately before IEF, and the slurry is poured into a horizontal trough. After 2–3 h IEF the respective protein fractions are applied directly on the IPG strip. The transfer of proteins into the IPG strip occurs electrophoretically with high efficiency. Figure 11.30 shows the separation results in a narrow pH interval without and with prefractionation.

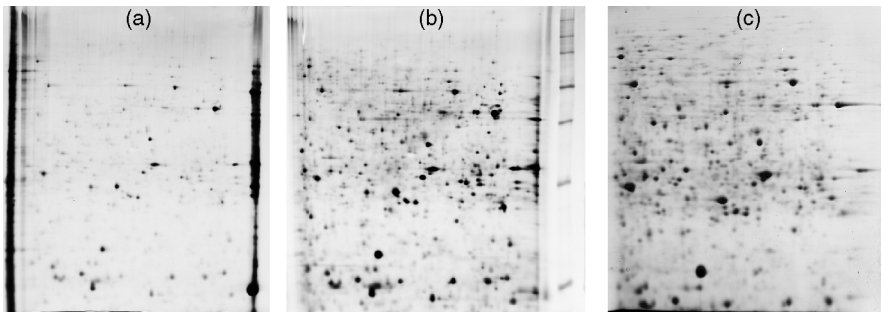


Figure 11.30 Two-dimensional electrophoreses in the narrow pH interval 6–7. Applied: (a) 250 µg complete protein mixture; (b) 250 µg protein prefractionated in a Sephadex-IEF gel; (c) 1 mg protein prefractionated in a Sephadex-IEF gel. (a) and (b): Silver stained; (c) Coomassie Blue staining. Source: Görg, A. *et al.* (2002) *Proteomics*, 2, 1652–1657. With permission, Copyright © 2002 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

11.6.3 First Dimension: IEF in IPG Strips

Dried film-backed gels with immobilized pH gradients are cut into narrow strips, which are rehydrated in urea solution containing the additional additives before isoelectric focusing. Because the pH gradient is fixed to the matrix, there are no edge effects during IEF. Different ways for sample application exist: For rehydration loading the urea-additive solution is mixed with the sample prior to rehydration. Then all proteins are distributed over the entire pH gradient and migrate in the electric field towards their pIs from different directions. This concept has the advantage that no protein precipitate is generated at a sample application point, and that it is possible to combine rehydration and IEF run in one methodical step; this simplifies the workflow and reduces the possibilities of errors. For cup loading (Figure 11.27) the sample is

applied on the already rehydrated gel at a defined pH value, mostly close to the anode. This variant is – with a few exceptions – highly recommended for basic pH intervals. Furthermore, there are protein mixtures that require this method for better spot patterns, particularly when they have to be separated in IPG strips longer than 13 cm. However, some sample cups do not allow loading of more than 100–150 µl sample volume, and protein precipitation at the sample application point can often be observed. Paper bridge loading allows the application of larger volumes. Here a filter card board strip is soaked in the sample solution and applied between one end of an IPG strip and the electrode pad. The electric current in such 3 mm wide and 0.5 mm thin IPG strips is in the micro-ampere range. Initially, only a very low field strength is applied to avoid aggregation and precipitation of proteins. At the end, IEF occurs at several thousand volts; the time is dependent on the strip length and the width (which is also the slope) of the pH gradient. The longest separation time – overnight – is needed for narrow intervals in 24 cm long IPG strips. The time is very critical for *basic pH gradients* – it should be as short as possible, because some proteins become instable at their pI in a very alkaline environment.

11.6.4 Second Dimension: SDS Polyacrylamide Gel Electrophoresis

It is very important to equilibrate the IPG strip prior to SDS electrophoresis for sufficiently long in a sufficient large volume of SDS sample buffer – first with a thiol reagent, then with an alkylation reagent. Alkylation with iodoacetamide prevents separation artifacts and converts the cysteines into a fully alkylated condition, which is also necessary for protein identification with mass spectrometry. A stacking gel is not needed, because the proteins are already pre-separated. When vertical gels are employed, the IPG strip is embedded with agarose to achieve gel continuity between the first and the second dimension. For horizontal systems no agarose is used, the IPG strip is simply placed into a groove, which is molded in the gel surface. In most cases a resolution power of a homogeneous gel with an acrylamide concentration of about 13% *T* is optimal. If the molecular sizes are widely spread and if there is a high amount of glycosylated proteins the separation result can be improved with pore gradient gels.

11.6.5 Detection and Identification of Proteins

The most important detection methods, imaging and image analysis, are described in Section 11.3.3. After 2D electrophoresis, usually the image analysis program has to perform qualitative and quantitative comparison of the highly complex spot patterns for the detection of newly expressed proteins or those that have disappeared, as well as regulated proteins.

The identification and characterization of these proteins is done with mass spectrometry. The protein spots are punched out from the gel, either manually or with spot picking robots. The proteins are digested into peptides with an endopeptidase, mostly trypsin. The peptides can be easily eluted from the gel plugs and submitted to mass spectrometry (Chapter 15).

Mass Spectrometry, Chapter 15

11.6.6 Difference Gel Electrophoresis (DIGE)

Difference gel electrophoresis considerably improves the reliability of qualitative and quantitative conclusions from 2D gel results. Prior to the separation the proteins of the different samples are labeled with different fluorescent tags. Then the samples are combined and separated together in the identical 2D gel: first in IEF and then in SDS electrophoresis. The dye molecules are designed so that the same proteins – coming from different samples – migrate to exactly the same positions in the gel. When the gel is scanned at different wavelengths, the – more or less – different 2D patterns are obtained in the respective emission channels. It is important that the excitation and emission wavelengths are sufficiently distant from each other to avoid cross-talks of signals.

Minimal Labeling For labeling of proteins at the ϵ-amino group of lysines the fluorescent tags contain a basic group to compensate for the loss of the positive charge due to blocking of the lysine. This prevents a shift of the pI. The labeling occurs via a *N*-hydroxysuccinimide (NHS) ester. The mass gain is the same for all dyes in order to prevent a shift in the molecule size axis.

Multiple labeling has to be avoided for several reasons. First of all the molecules would become very hydrophobic and not stay in solution. The 2D pattern should be comparable with results obtained with unlabeled samples. Furthermore, downstream analysis with mass spectrometry and Western blotting would be disturbed. Therefore, the dye to protein ratio is kept very low (minimal labeling). When only 3–5% of the total amount of all proteins carries a label, multiply labeled proteins will not be detected because of their statistically low concentration. The fluorescent dyes used still emit a strong signal, reaching a sensitivity that is almost comparable to silver staining. The 2D pattern is not changed compared to those obtained with unlabeled proteins. For downstream analysis, 95% of the proteins remain without a label. For low molecular weight proteins smaller than about 20 kDa the gain in molecular weight by about 460 Da can cause a slightly slower migration of the labeled proteins in the second dimension. To be on the safe side, gels for spot picking should also be post-stained to ensure accurate punching of the unlabeled small proteins spots. For lysine labeling three to four different dyes are available.

Saturation Labeling Labeling of cysteines occurs via a maleimide group. This labeling method is charge neutral; there is no shift of the isoelectric point. In this approach all available cysteines are labeled. Therefore, the molecular weights become modified depending on the number of cysteines in a protein. Prior to labeling the disulfide bridges have to be cleaved with a thiol reagent, like dithiothreitol, or with Tris-(hydroxycarboxyl)phosphine. Also with this labeling the differently labeled proteins from different samples migrate to the same positions. But the patterns differ considerably from those produced by unlabeled or minimal labeled proteins, because there are multiple labels depending on the cysteine contents of each protein. This technique offers the detection of proteins far above the limit of detection of silver staining – but only for cysteine-containing proteins. Proteins without cysteine are not detected. For cysteine labeling only two different dyes are available.

In some cases it could be observed that some proteins show preferred labeling with one of the dyes. To prevent quantitative errors originating from this effect the application of a systematic dye swap between wildtype and mutants is advised – and to analyze even numbers of sample replicates instead of odds.

Internal Standard DIGE offers the unique possibility to include an internal standard for each protein contained in the samples. This feature increases the statistical certainty of the results considerably – and it reduces at the same time the need for technical replicates. For this, aliquots of each sample are taken and mixed together. This mixture is labeled with one of the available fluorescent tags.

The labeled standard is combined with the labeled sample(s) and co-electrophoresed with the samples (Figure 11.31). To evaluate the protein patterns all measured spot positions and volumes are related to the respective standard spot and normalized. The results are relative quantitative values for increases or decreases of protein expression levels in the compared samples down to a minimum value of 5% at a statistical confidence of more than 95%. Such results are by far not achievable with conventional staining techniques, not even with five technical replicates.

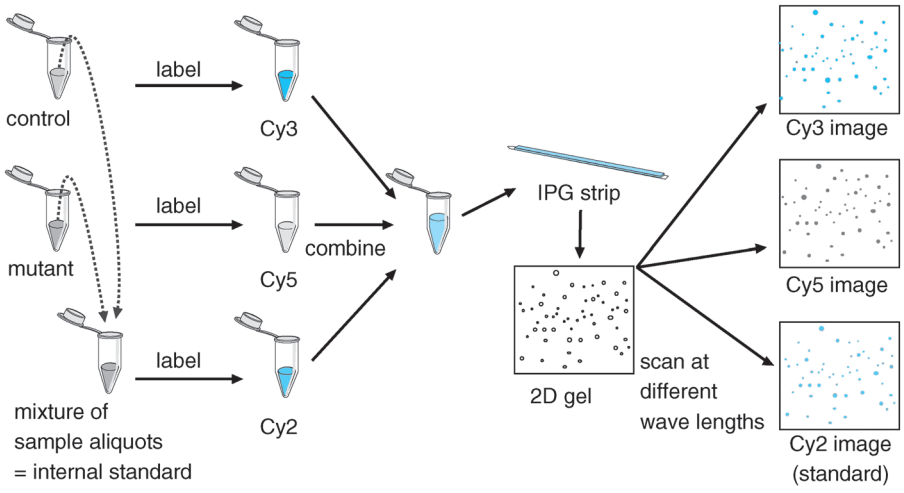


Figure 11.31 Principle of DIGE. Cy2, Cy3, and Cy5 are fluorescent dyes, which bind to proteins. Example for two different samples. When more samples are analyzed, more gels are needed, but each gel also contains the internal standard, which is created by a mixture from aliquots taken from each sample of the experiment.

11.7 Electroblotting

After gel electrophoresis and focusing the proteins are enclosed in the matrix and are not accessible for antibodies or lectins for specific immunological or affinity detection methods. Furthermore, there have always been attempts to acquire the separated protein fractions in liquid solution for subsequent analysis, for example, via diffusion, electroelution, or extraction with acids or organic solvents. Instead of extracting the proteins into a solution, two research groups – J. Renart *et al.* and H. Towbin *et al.* – both described at the same time, in 1979, an alternative approach to transfer the electrophoretically separated proteins from the polyacrylamide matrix onto a nitrocellulose membrane with the help of an electric field, where they become immobilized. This method has established itself as Western blotting; this technique allows one to detect the separated proteins (e.g., antigens, glycoproteins, or enzymes) carrying specific binding properties with antibodies, lectins, or enzyme substrates *directly* on the membrane. Prior to probing with ligands the non-occupied binding-sites on the membrane surface have to be blocked with inert macromolecular substances, such as with bovine serum albumin, fish gelatin, skin milk powder, or detergents like Tween 20. The blocking substance has to be specifically selected to avoid cross reactions with the antigens.

Western Blotting Chapter 5.3.3

11.7.1 Blot Systems

For the electrophoretic transfer two different systems are employed, tank blotting and semidry blotting.

Tank Blotting The standard equipment for tank blotting is a vertical buffer tank containing meandering platinum electrode wires attached to two sidewalls (Figure 11.32). The slab gel and the membrane are laid between filter papers and thin sponges, which are soaked with buffer, and placed into a grid cassette. During packing of the cassettes air bubbles must be avoided. The cassettes are inserted vertically into the buffer tank. Usually, the transfers are performed with a constant voltage of 50 V in order to execute a continuous force on the charge carriers in a constant electric field. The starting value of the current lies at 500 mA or higher, depending on the size of the tank and the molarity of the blotting buffers, and decreases during the transfer due to a continuously increasing ohmic resistance. Under these conditions efficient cooling is needed, which can be achieved with a vertical heat exchanger and an efficient buffer circulation.

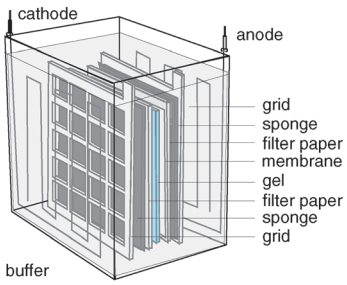


Figure 11.32 Blotting tank for electrophoretic transfers of separated proteins on immobilizing membranes. The meandering electrode wires are attached to the front and back wall. The gel and the membrane are clamped between filter papers, sponges, and grids.

Semidry Blotting Semidry blotting appliances, which were been described in 1984 by Kyhse-Andersen, consist of two horizontal electrode plates, between which the blot sandwich (slab gel, membrane, filter papers soaked with buffer on both sides) is placed (Figure 11.33). Compared to tank blotting the setup is simpler, because no cassettes and only small buffer volumes are used. Air bubbles are removed with a roller. The small buffer volume also has the advantage that the

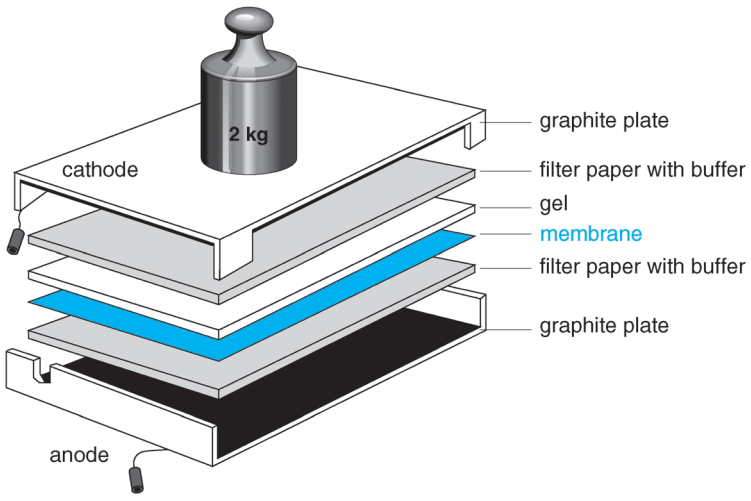


Figure 11.33 Semidry blotting. Gel and membrane are placed between filter paper soaked in buffer.

proteins experience less exposure to reactive impurities than in tank blotting. There are various plate electrode materials on the market, which differ in electric conductivity, stability against anodal oxidation reactions, and extreme pH values, for example, very pure graphite, glassy carbon, graphite in a plastic matrix, platinized metal sheets, and conducting plastic. In most cases semidry blots are carried out with constant current (e.g., with 0.8 mA per cm^2 blot area), which causes a very low voltage (<5 V) at the beginning. During the transfer the voltage increases with the increasing ohmic resistance, dependent on the type of blotting buffer, the amount of blotting buffer, the gel thickness, and the electrode plate material. Because of the high conductivity of the electrode plates and the low electric power generated, cooling of the semidry apparatus is not necessary. The electrochemical reaction of water produces a pH gradient from about pH 12 at the cathode ($4H_2O + 4e^- \rightarrow 2H_2 + 4OH^-$) to about pH 2 at the anode ($6H_2O \rightarrow O_2 + 4H_3O^+ + 4e^-$), as well as continuous gas development of the reaction products. The gas pushes the blot sandwich apart, and causes an irregular distribution of electric field strength. Weighting the upper blotting plate with a ca. 2 kg weight assures an even and reproducible distribution of electric field (Figure 11.33). The stability of the electrode material differs: All graphite electrodes and graphite-plastics are attacked by nascent oxygen under production of CO_2 – faster or slower, depending on the quality. Platinum or platinized electrodes are practically inert. During recent years semidry blotting has become more accepted than tank blotting for two reasons: First, handling is easier and, second, systematic comparisons in the literature show that protein transfers are more homogeneous and more efficient at a shorter transfer time. Furthermore, immunodetection and protein staining are more sensitive, because the proteins migrate shorter distances through the membrane and remain on the upper surface.

11.7.2 Transfer Buffers

The mostly used transfer buffer consists of Tris-glycine pH 8.3 and 20% (v/v) methanol, which can be replaced without any problems by ethanol or isopropanol. The alcohol has two functions: it improves the binding properties of the membrane and it prevents swelling of the gel slabs. Swelling of gels can cause spatial smearing of protein zones. Alternatively, CAPS buffer adjusted to pH 10.5 with NaOH is often used, because its high pH value can increase the transfer efficiency. For semidry blotting the discontinuous buffer, described by Kyhse-Andersen exhibits a better transfer efficiency than the Tris-glycine buffer. For electroblotting of IEF gels usually 1% (v/v) acetic acid is used; the transfer direction is converted towards the cathode. However, for the transfer of proteins from isoelectric focusing gels, particularly when they contain urea, capillary blotting (see Figure 6.22) or pressure blotting methods are superior to electroblotting, because at their pIs the proteins have a net charge of zero.

11.7.3 Blot Membranes

For protein blotting either nitrocellulose or poly(vinylidene fluoride) (PVDF) membranes are used with 0.2 and 0.45 µm pore diameters. PVDF membranes are chemically inert and possess superior mechanical properties and higher binding capacities than nitrocellulose, but they are hydrophobic and have to be pre-wetted with 100% alcohol before use. Nitrocellulose binds low molecular weight proteins and peptides more efficiently than PVDF. For detection with fluorescent labeled secondary antibodies, blot membranes with out or low autofluorescence have to be selected. For experiments with subsequent protein-chemical analyses, for example, amino acid sequence analysis, only PVDF membranes are employed due to their high chemical stability.

Sequencing Yield, Section 14.1.3

MALDI-MS, Chapter 15

Further Reading

Andrews, A.T. (1986) *Electrophoresis, Theory, Techniques and Biochemical and Clinical Applications*, Clarendon Press, Oxford.

Celis, J.E. (ed.) (1998) *Cell Biology. A Laboratory Handbook*, Academic Press, San Diego.

Righetti, P.G. (1990) *Immobilized pH Gradients: Theory and Methodology* (eds R.H. Burdon and P. H. Knippenberg), Elsevier, Amsterdam.

Rothe, G.M. (1994) *Electrophoresis of Enzymes. Laboratory Methods*, Springer, Heidelberg.
Simpson, R.J. (ed.) (2004) *Purifying Proteins for Proteomics*, Cold Spring Harbor Laboratory Press, New York.
Westermeier, R. (2016) *Electrophoresis in Practice*, Wiley-VCH Verlag GmbH, Weinheim.
Westermeier, R, Naven, T, and Höpker, H.R. (2008) *Proteomics in Practice*, Wiley-VCH Verlag GmbH, Weinheim.

Capillary Electrophoresis

12

Philippe Schmitt-Kopplin,[1,2] Constanze Müller,[1] Jenny Uhl,[1] and Arthur Wayne Garisson[3]

[1]Helmholtz Zentrum München - German Research Center for Environmental Health, Analytical BioGeoChemistry, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany
[2]Technische Universität München, Chair of Analytical Food Chemistry, Alte Akademie 10, 85354 Freising-Weihenstephan, Germany
[3]U.S. Environmental Protection Agency, National Exposure Research Laboratory, Ecosystems Research Division, Athens, Georgia, USA

Capillary electrophoresis (CE) has found a perfect place next to chromatography in the world of separation sciences through the versatility of the possible separation principles and applications. The method became well-known very early on as *the* analysis technique in molecular biology: many thousands of instruments are used worldwide for DNA sequencing and genotyping. Furthermore, CE methods became standardized for the analysis of small ions and small molecules to (bio)polymers and proved complementary for the analysis of proteins and peptides as well as for classical technologies such as 2D gel electrophoresis and HPLC. Especially for difficult applications (recombinant proteins, monoclonal antibodies, chiral separations), the high resolution and selectivity of CE enabled high quality separations. In the field of small molecules it still shows unique potential for highly polar and charged compounds that are difficult to separate with chromatography. Besides important applications in the bioanalysis of biomacromolecules (peptides, proteins, saccharides, DNA, RNA, etc.), CE has also found valuable analytical applications in the fields of pharmaceuticals (chiral molecules, vitamins), forensics (DNA, drugs of abuse), cell biology (analysis of whole cells, bacteria, viruses), environmental chemistry (pesticides, tensides), metabolomics (in medical and nutritional sciences), and the chemical industry (polymers, dyes, detergents). One of the attractions of CE is its use of a minimal amount of sample (submicroliter), low solvent consumption, short analysis times, and high resolution. In addition, it is an orthogonal and thus complementary separation approach to HPLC and GC with similar detection and evaluation procedures.

12.1 Historical Overview

The fundamentals for the rapid development of CE in the last few decades were established over the last century. The principles of electrophoresis as related to the mobility of charged particles in an electrical field were first described in detail by Kohlrausch (1897). Tiselius developed electrophoresis as an analytical method for proteins and was honored with the Nobel Prize in Chemistry 1948 for his achievements. With the introduction of anticonvection media (paper, polyacrylamide, and agarose gels), electrophoresis based tools became indispensable for biochemistry. These techniques, involving gel polymerization, colorization, and densitometry are, however, very work intensive and, in addition, some interactions between the sample and the gel matrix cannot be excluded. The first approaches thus led to the direct separation of the analytes in the buffer solution with photometric detection. In thin tubes, thermal convection is minimized because of good thermal conductance at high surface to i.d. ratios. Thus, a higher voltage can be applied and shorter separation times achieved. Hjertén (1958) showed the first separation in open tubes of 3 mm i.d. The convection was minimized by rotation around the main tube axis and the analyte zone was localized with UV-detection. The methylcellulose coating of the glass walls reduced the electroosmotic flow. Further reduction of the i.d. to

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

200 µm in a 1 m length capillary by Virtanen enabled the separation of alkali metal ions. Everaerts (1970) used 500 µm i.d. Teflon tubing to separate organic acids by isotachophoresis with online thermometric detection. With the introduction of conductivity and UV-detection to capillary electrophoresis by Mikkers (1979), high resolution separation/detection could be achieved. The real birth of capillary electrophoresis was in 1981 when Jorgenson used open tubular quartz capillaries with 75 µm i.d. under a separation potential of 30 000 V to separate derivatized amino acids and peptides, with fluorescence detection. The efficiency approached theoretical limits and interest in capillary electrophoresis increased tremendously. The number of publications at that time increased as well, reflecting many technical developments, such as better reproducibility of the glass capillaries, smaller capillary i.d., and miniaturization. Capillary electrophoresis profited from the development of HPLC in terms of automation and improved software development, and further developments over the years enabled submicrometer instrumentation levels (microchip-CE technology), enabling the analysis of single cells and even single molecules.

12.2 Capillary Electrophoresis Setup

Instrumental Setup Compared to classical chromatography (LC, GC), CE is a relatively simple technology. The main constituents are shown in Figure 12.1 and involve the separation capillary, voltage unit, electrodes, buffer reservoirs, and on-column detection unit. The system is computerized, thermally regulated (capillary and samples), and enables hydrodynamic or electrokinetic injections.

Voltage Unit CE instruments are equipped with voltage units that deliver ±30 kV. They should be able to work at constant voltage and constant current. Higher voltages are in principle possible but can lead to electric discharges caused by air humidity in particular with shorter capillaries. The application determines whether detection is to be at the anode or cathode and thus the polarity must be changeable. The voltage is led to the platinum electrodes that are plunged into the buffer/sample vials.

Capillary The most used capillary material is fused silica (amorphous quartz). Additionally, borosilicate, PTFE (Teflon, polytetrafluoroethylene) or PEEK (polyetheretherketone) that were used earlier for isotachophoresis are available for specific applications. In addition to mechanical properties, one of the main column issues is UV-transparency for optical detection, which limits use to only a few materials. The need for small inner diameters for the most efficient heat dissipation, to enable the high resolution of CE, was discovered by Tiselius. However, the technical realization only was possible in the 1980s, allowing a breakthrough in the applications of CE. Inner diameters are usually chosen in the range 50–100 µm. Low i.d. columns may become

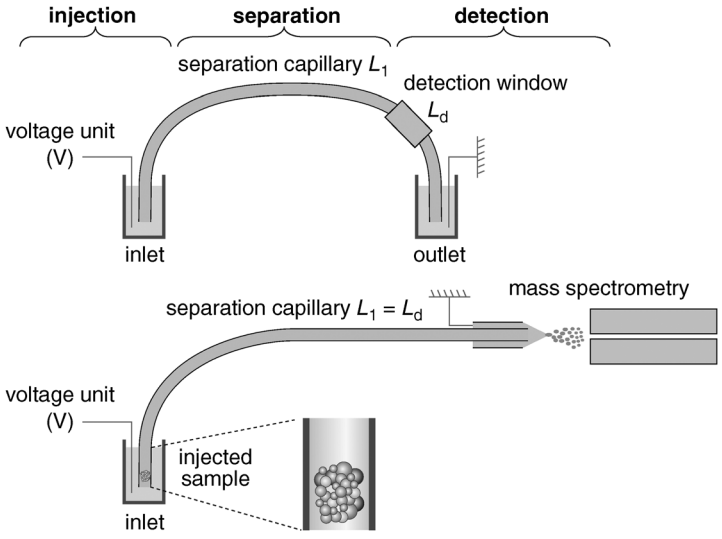


Figure 12.1 Schematic view of capillary electrophoresis.

blocked when small particles are present in the sample. To enhance mechanical stability a 10 µm polyimide layer is coated onto the outer surface of the capillaries. To enable UV-Vis detection, an optical window can be made in the column by burning with a flame, mechanically scraping with a scalpel, or chemically dissolving with concentrated sulfuric acid; great care has to be taken as this window part of the separation capillary is especially fragile. Further miniaturization of CE systems has led to microchip electrophoresis (MCE) using glass, thermoplastic polymers like polydimethylsiloxane (PDMS), or poly(methylmethacrylate) (PMMA) platforms.

12.3 Basic Principles of Capillary Electrophoresis

12.3.1 Sample Injection

The injection volume needs to be kept as small as possible to guarantee the high separation efficiency of capillary electrophoresis by avoiding additional band broadening. The variance of the additional peak broadening σ^2_{inj} is dependent of the length (l) of the injection plug as:

$$\sigma^2_{inj} = \frac{l^2}{12} \tag{12.1}$$

The injection reproducibility of these small sample volumes (only a few nanoliters for a column volume of about a microliter) is the main issue for the reproducible use of capillary electrophoresis in routine analysis. Two injection modes are usually available:

- hydrodynamic injection,
- electrokinetic injection.

Hydrodynamic Injection This is the most used sample injection mode in CE. The injected volume can be described by the Hagen–Poiseuille equation (with the pressure difference ΔP in Pa, the capillary inner diameter d, time t, viscosity η, and capillary length L):

$$V_{inj} = \frac{\Delta p\left(d^4 \pi t\right)}{128 \eta L} 10^3 \tag{12.2}$$

After plunging the capillary into the sample liquid, various modes of establishing the pressure difference can be applied:

- applying pressure on the inlet side,
- applying a vacuum on the outlet side,
- by gravity, that is, placing the inlet at a higher position than the outlet.

Oversized injection volumes (sample zone >10% of the capillary volume) result in signal distortions and severe band broadening. Thus, especially when using methods from the literature with different capillary lengths or i.d., one should always adapt the injection parameters. For most methods, electrokinetic injection is the most popular as it is easily automated with pressure corrections that allow standard deviations in injection volume of around 1%.

Sample Stacking In sample stacking one takes advantage of the fact that the electric field along an electrolyte filled capillary is inversely proportional to the conductivity of the electrolyte at this position. Thus the ions of a zone of lower conductivity attain a higher speed until they reach a zone of higher conductivity where they are slowed down. The triggering of various conductivity buffer zones can thus be used to guide the concentration of ions to selected capillary positions and sharpen the sample zone for better separation efficiency and sensitivity. It is thus important that the conductivity of the sample is always an order of magnitude lower than that of the buffer.

Electrokinetic Injection In this mode, sample injection occurs after applying a voltage between the capillary end and the sample, creating a difference of potential that enables an electrophoretic and electroosmotic transport of the sample into the capillary. Compared with hydrodynamic

injection, anions and cations can be discriminated as a function of the applied voltage. The injected amount of sample is proportional to the applied voltage and is increased for ions of higher mobility (small ions, multiple charges). Additionally, the injected amount of analytes depends on the sample matrix, the higher the share and mobility of the matrix is, the smaller is the amount of analytes; if the sample is dissolved in water the ions are strongly concentrated in the capillary. Electrokinetic injections are highly reproducible and technically easy to realize. However, different matrix constitutions may lead to variation in injected amounts (up to a factor of 100), necessitating the use of internal standards. Knowledge of the relative ion mobility can also be used for selective injection (sample pH dependent ion injections as in CEkF – capillary electrokinetic fractionation). Most of the time electrokinetic injection is used with filled capillaries such as in capillary gel electrophoresis (CGE) or capillary electrokinetic chromatography (CEC) where hydrodynamic injections are not possible.

12.3.2 The Engine: Electroosmotic Flow (EOF)

During the analysis of strongly positive charged particles electrostatic interactions with negatively charged silanol groups (on the capillary surface) can drastically impair the separation (signal broadening). In these cases the separation can either be enhanced by chemical modification of the silanol groups or by adsorption of polymers or positively charged detergents to the surface (dynamic coating).

In capillary electrophoresis the use of fused silica capillaries enables an electroosmotic flow (EOF), leading to a net flux of buffer in the capillary; the velocity of the ions as measured at the detector is thus due to a combination of the effective mobility of the ions and the velocity of the buffer as expressed in Figure 12.2. The zeta-potential and thus the EOF is directly dependent on the pH-controlled association of the silanol groups at the capillary surface, and thus on the pH and ionic strength of the buffer solution (Figure 12.3). At alkaline pH the EOF is higher than the mobility of the ions, enabling the transport of anions towards the cathode; at low pH the EOF can be completely suppressed. Thus, depending on the pH of the separation buffer, anions, cations, and neutral molecules can be separated.

The decline in potential in the diffuse double layer at the column inner surface is dependent on the ionic strength of the buffer solution; the higher the ionic strength, the stronger the potential difference and the lower the diffuse double layer. The EOF thus declines with the increase in ionic strength.

Compared to the usual parabolic hydrodynamic flow, the profile of the EOF is quite laminar (Figure 12.4), so the fluid velocity is equal over all the capillary width (except at the double layer where the EOF varies from zero to the maximum value). This velocity homogeneity helps to avoid peak broadening such as observed in chromatographic separations, so that separations can be performed even at high EOF without peak shape alterations.

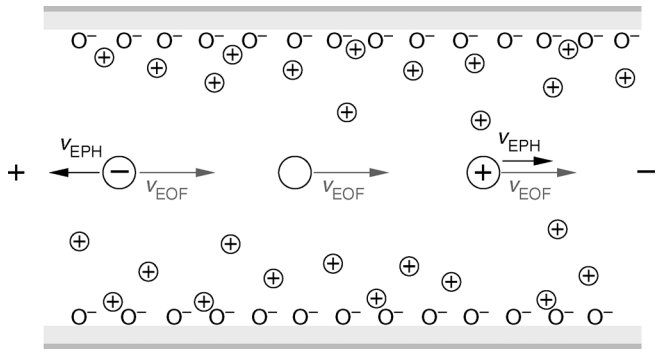
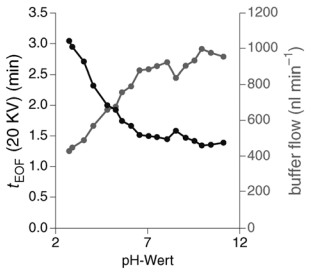


Figure 12.2 Principle of zone electrophoresis: migration of ions and neutral particles within the capillary according to electroosmotic flow (v_{EPH} = electrophoretic velocity, v_{EOF} = electroosmotic velocity).



t_{EOF} min	ID 100 µm nl min^{-1}	ID 75 µm nl min^{-1}	ID 50 µm nl min^{-1}	ID 20 µm nl min^{-1}
2.0	1964	1105	491	79
2.5	1571	884	393	63
3.0	1309	736	327	52
3.5	1122	631	280	45
4.0	982	552	245	39
4.5	873	491	218	35
5.0	785	442	196	31
5.5	714	402	178	29
6.0	655	368	164	26
6.5	604	340	151	24

Figure 12.3 The pH dependence of electroosmotic flow at constant ionic strength. The buffer flow rate inside the capillary is changing in correspondence to the pH and thus the time t_{EOF}, in which uncharged particles reach the detector. Source: Schmitt-Kopplin, Ph. and Frommberger, M. (2003) *Electrophoresis*, **24**, 3837–3867. With permission, Copyright © 2003 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

Dynamic coating of the capillary walls with polymers (based on poly(ethylene glycol)s, celluloses or poly(vinyl alcohol)s) enables a reduction of the EOF (see Equation 12.10 in Section 12.4.8) because of the increase of viscosity at the double layer. Coating with positively charged detergents (cetyl(trimethyl)ammonium bromide – CTAB) enables inversion of the EOF. Furthermore, chemical modification of the silanol groups on the capillary surface can lead to hydrophilic or hydrophobic coatings. The main problem with this approach is the possible lack of stability to hydrolysis over long time periods. Dynamic replaceable coatings are always preferred for better reproducibility.

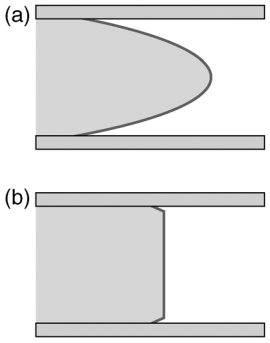


Figure 12.4 Comparison of different capillary flow profiles: (a) hydrodynamic parabolic-shaped profile, (b) EOF-profile.

12.3.3 Joule Heating

During an electrophoretic separation, heat is developed from the electric flux in the leading electrolyte system. Heat dissipation can occur through the capillary walls, leading to a radial temperature gradient through the capillary section (Figure 12.5). To enable maximum separation efficiency the temperature gradient must be kept as small as possible; this is achieved even with high currents by keeping the capillary i.d. smaller than 100 µm and actively cooling the capillary by immersion in an air or liquid bath. Under optimized conditions the temperature differences through the gradient are less than 1 °C.

12.3.4 Detection Methods

For detection of the analytes after their separation by capillary electrophoresis, the following detection methods are possible:

- absorption based detection:
  - UV-detection
  - diode array detection;
- fluorescence detection:
  - laser induced fluorescence;
- mass spectrometric detection;
- electrochemical detection;
- radioisotopic detection;
- conductivity detection;
- light refraction detection;
- magnetic resonance spectrometry detection.

Most commercial instruments have adsorption based detectors and can optionally be equipped with more specific detection modes. Spectroscopic detection (UV, fluorescence) occurs generally through the capillary walls, and so the external polyimide coating on the fused silica capillary needs to be removed (burning a millimeter-wide window or using concentrated sulfuric acid).

Due to the small i.d. of the capillary the light path is short, limiting the sensitivity of the CE technique. For example, for peptides concentrations of 100 ng l^{-1} are usual. However,

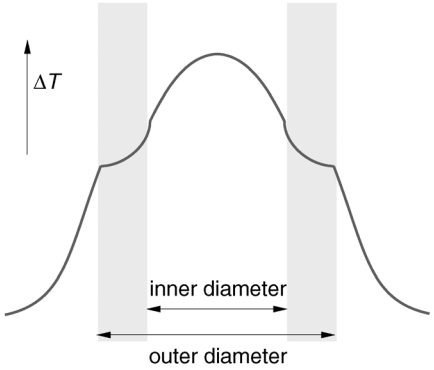


Figure 12.5 Temperature gradient within a capillary cross section.

fluorescence detection could easily improve the sensitivity by a factor of 1000; the only problem is the need for selective derivatization, with its known limitations, as in chromatography. Coupling of CE to mass spectrometers is possible via electrospray (ESI), photo- or laser-ionization (APPI, APLI), or chemical ionization (APCI). Online matrix-assisted laser desorption ionization (MALDI) has also been reported for separation in proteomics applications. The other listed detection possibilities are more specific to the needs of some extreme applications.

UV-Detection Variable wavelength and diode array detectors are the most commonly used detectors in capillary electrophoresis. The detection window should be smaller than the sample zone (typically 1–5 mm) to avoid poor resolution at the detector and is usually around 1 mm long; for fast separations, this can be a limiting factor for separation efficiency. With absorption detectors the limiting factor is given by the Lambert–Beer law, since the light path corresponds to the capillary i.d. Some attempts to increase the light path were bubble type widening at the detector window or *z*-type coaxial elongation of the path, but these were only applicable for well-separated analytes in specific routine applications because of losses in separation efficiency. The best solution is an increase in concentration of the sample with choice of optimum wavelength for each analyte by using a diode array detector. Here automatic peak recognition can be used and the CE systems are programmable, as in liquid chromatography.

UV/VIS Spectroscopy, Section 7.1-7.2

Fluorescence Detection Fluorescence excitation is possible with deuterium, pulsed xenon, or laser sources. However, to concentrate the needed energy on the small volume of the detection chamber laser induced fluorescence (LIF) is the best solution. Compared to absorption-based detection, where the ratio of the incident to absorbed light energy is important, in fluorescence detection the signal intensity is directly proportional to the incident excitation energy. Thus, fluorescence detection is highly structure selective and can be extremely sensitive; fluorescein can be detected at 10^{-12} molar. Compounds with native fluorescence (especially tryptophan containing peptides and proteins) can be directly exited with a UV-laser and many possibilities were developed based on selective fluorescent dyes. Some of the combinations of laser types, excitation wavelengths, dyes, and applications are summarized in Table 12.1.

Fluorescence Spectroscopy, Section 7.2

Mass Spectrometry Detection The high separation efficiency of CE combined with the mass selectivity and accurate mass detection of MS is a very promising technology. The online coupling of CE to MS is possible in various setups (Figure 12.6). In the setup without sheath flow the electrical contact is obtained either through a liquid-junction or a conducting capillary end (microelectrode or metal coated capillary). The most robust setup, which is also commercially available, is coupling via a coaxial solution sheath flow (2–10 µl min^{-1}), which delivers counter-ions for the CE separation, makes the electrical contact, and enhances the spray stability. One

Mass Spectrometry, Chapter 15

Table 12.1 Laser for fluorescence excitation.

Source of energy	λ_{EM} (nm)	Fluorescence dyes (examples)	Applications
Argon ion laser	488	FITC (fluorescein isothiocyanate)	Peptides, DNA
		NBD-F (4-fluoro-7-nitrobenzofuran)	Amino acids, peptides
		APTS (1-aminopyrine-3,6,8-thiosulfuric acid)	Oligosaccharides
		FQ (5-furoyl quinolone-3-carboxaldehyde)	Proteins
Helium–neon laser	544	SYPRO Red	Proteins
		Merocyanine 540	Proteins
Helium–cadmium laser	325	Dns-Cl (dansyl chloride)	Amino acids
		ANTS (8-aminonaphthalene-1,3,6-trisulfonic acid)	Oligosaccharides
Diode laser	635	Cy 5 (cyanine dye)	DNA, antibodies
	670	Dicarboxylic cyanine	Amino acids, peptides
Solid state laser	355	Fluorescamine	Amino acids, peptides

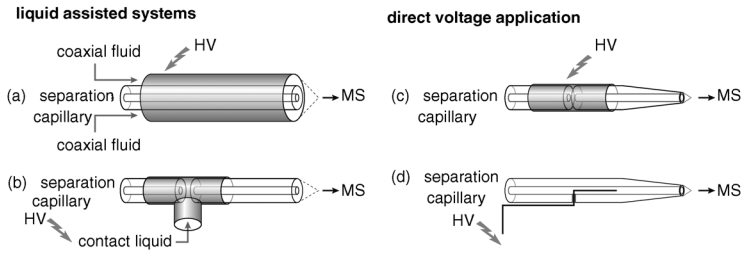


Figure 12.6 Types of ESI interfaces. (a) system with coaxial fluid (sheath liquid): voltage application on outer metal tube and power relay via contact fluid on the end of separation capillary (typical diameter approximately 1 mm); (b) liquid bypass (liquid junction): voltage application distal on the end of separation capillary; (c) sheathless system (voltage application distal on the end of separation capillary or on conductive coating of ionization needle); (d) voltage application on platinic in-column electrode. Source: Schmitt-Kopplin, Ph. and Frommberger, M. (2003) *Electrophoresis*, **24**, 3837–3867.

major limitation in using MS-detection with CE is the need for MS-compatible buffer systems, that is, volatile buffers (usually ammonium salts of formate or acetate), which often lowers the sensitivity of electrospray mass spectrometry. However, the acquisition of exact mass with the newest generation of mass spectrometers makes this tool almost indispensable in modern bioanalytical laboratories for obtaining the highest selectivity and accuracy.

12.4 Capillary Electrophoresis Methods

The various capillary electrophoresis methods are based on a variety of buffer systems and hardware setups, as summarized in Table 12.2.

12.4.1 Capillary Zone Electrophoresis (CZE)

Capillary zone electrophoresis (CZE) is the basic mode of capillary electrophoresis; separation is based only on the differential mobility of the analytes due to charge and size. The capillary is homogeneously filled with an electrolyte system that has the role of maintaining a constant current at a constant pH. Figure 12.7 shows the separation, schematically, and Figure 12.8 illustrates the experimental pH-dependent electropherograms of pesticide metabolites as cations and anions. The simplest optimization of a separation is through control of the pH of the separation buffer system. As the effective mobilities of ions are dependent on their dissociation degree, the best separation of closely related analytes is at the pH corresponding to their pK_a ranges.

Table 12.2 Capillary electrophoresis separation techniques.

Separation technique		Separation according to differences in	Applications
Capillary zone electrophoresis	CZE	Size/charge (mobility)	Small ions, peptides, proteins
Isotachophoresis	ITP	Size/charge (mobility)	Small ions, proteins
Affinity capillary electrophoresis	ACE	Size/charge (mobility)	Ligand interaction
Non-aqueous capillary electrophoresis	NACE	Size/charge (mobility)	Small nonpolar ions
Micellar electrokinetic chromatography	MEKC/MECC	Hydrophobicity/charge	Uncharged particles
Capillary gel electrophoresis	CGE	Size	Proteins, DNA
Capillary electrochromatography	CEC	Chromatographic retardation	Small ions, uncharged particles
Isoelectric focusing	CIEF	Charge (isoelectric point)	Proteins

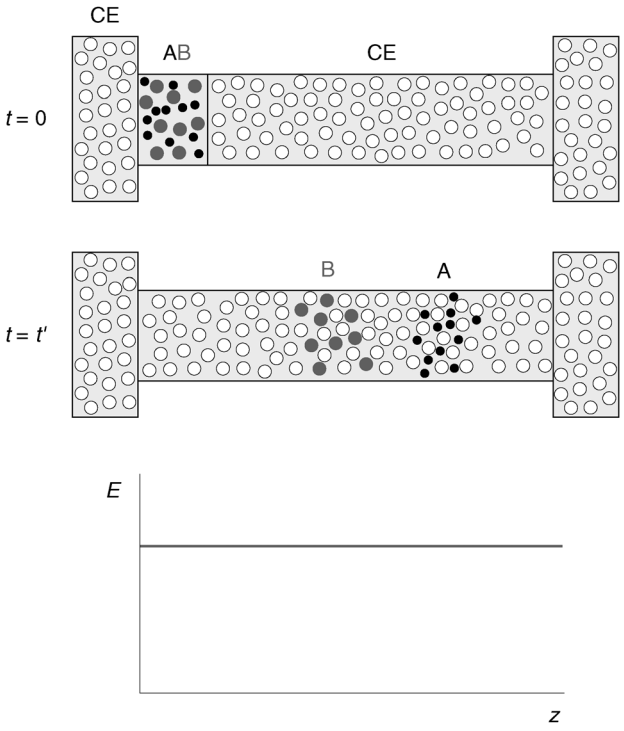


Figure 12.7 Principle of zone electrophoresis. Capillary is completely filled with electrolyte (CE). Ideally, the consistent field intensity is not influenced by sample ions. Sample ions A and B migrate at different speeds due to varying mobility. Diffusion leads to zone broadening.

Peak Broadening In CZE, analyte diffusion is ideally only longitudinal and is minimized. The number of theoretical plates is only dependent on the voltage applied across the capillary (and not on the capillary length) and the charge of the analyte (the only substance specific parameter). Practically, the maximum number of theoretical plates is never reached as many other parameters contribute to peak broadening, including temperature gradients, sorption to the capillary wall, and electrodispersion interferences during injection (hydrodynamic and electrokinetic) due to differences in mobility between the analyte and the background electrolyte ions.

Different peak shapes are often in an electropherogram: leading, symmetric and tailing. The reasons are not in the adsorption of analytes to the capillary wall, but only in the different mobilities. This problem can be solved by adapting the mobility of the buffer ions to those of the analyte or by the usage of lower sample concentration or higher buffer concentration.

Electrodispersion The peak broadening in CZE is only due to longitudinal diffusion when the electric field is constant over all the separation capillary and is not disturbed by conductivity heterogeneity. This may only be the case when the mobility of the analyte and buffer ions are very close or when the concentration of the analyte ions is much lower than that of the buffer electrolyte (factor of 100). In all other cases, additional peak broadening occurs because the electric field in the sample zone is different from the one in the surrounding buffer system. The resulting profile is not Gaussian but shows *leading* and *tailing*.

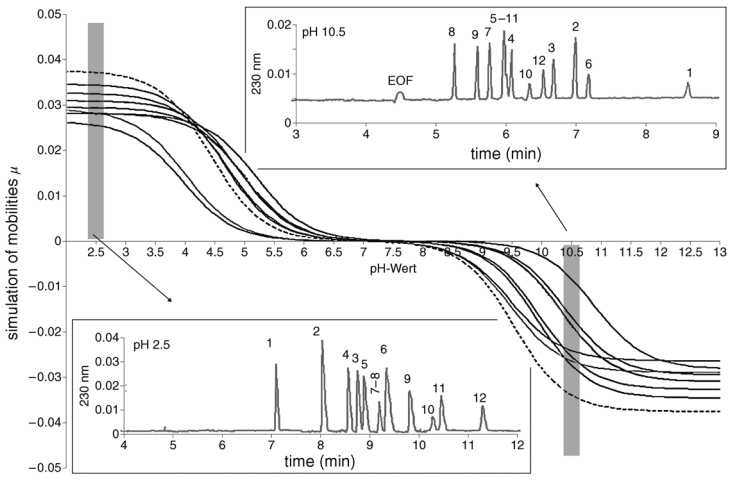


Figure 12.8 Separation of *s*-triazine derivatives as cations (pH 2.5) and anions (pH 10.5) by zone electrophoresis: simulation of mobilities at different pH values enables the estimation of separation as a function of the p*K* values of analytes. Source: Schmitt-Kopplin, Ph., Poiger, T., Simon, R., Garrison, A.W., Freitag, D., and Kettrup, A. (1997) *Anal. Chem.*, **69**, 2559–2566. With permission, Copyright © 1997, American Chemical Society.

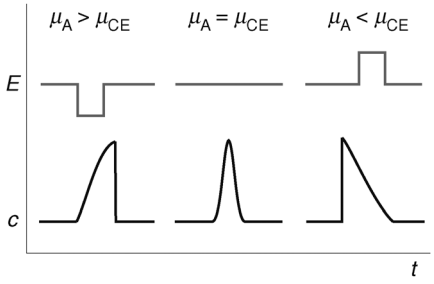


Figure 12.9 Distribution of concentrations and progression of field intensities at variable mobilities of samples and buffers (μ_A mobility of sample, μ_{CE} mobility of buffer).

When the mobility (μ_A) of the sample ions is lower than that of the buffer ions with the same charge (μ_{CE}), the sample zone has a lower conductivity and higher electric field than the buffer system. The front of the peak is thus sharper than the more diffuse *tailing* (diffusion into the separation buffer with lower electric field induces a lower mobility). In the case of a higher mobility of the sample ions the phenomenon is inverted and the profile is *leading* (Figure 12.9).

A massive peak broadening also occurs when the salt concentration of the sample is very high, that is, when the conductivity of the sample is much higher than that of the buffer. In CZE the sample thus should contain as little salt as possible, especially when higher volumes need to be injected. Ideally, one would have the sample in pure water; thereby, the initial sample zone is compressed (stacking). By using this way of in-line sample concentration, the volume of the sample can be increased by a factor five to ten to increase sensitivity.

Optimization of Separation Different parameters can be tuned to optimize the resolution of separations:

- pH-value of the buffer
- ionic strength of the buffer
- temperature
- capillary coating
- buffer additives.

pH-Value of the Buffer The buffer pH has a direct effect on (i) the partial charge of the acid/base analytes and, thus, on their electrophoretic mobility and (ii) on the electroosmotic flow (pH dependent zeta potential and increasing EOF with increasing pH). A higher pH leads to shorter analysis times (cationic mode) but also to lower resolution of the ions with $\mu_i < \mu_{EOF}/2$. These effects can nicely be modeled for migration predictions (Figure 12.8).

Ionic Strength The buffer ionic strength has a direct effect on the EOF and on mobility of the ions. Higher ionic strength has the advantage of lower electrodispersion with higher concentrated samples (sharper signals). Electrostatic interactions of the analytes (i.e., proteins) with the capillary wall can be reduced with increasing ionic strength. However, high ionic strength and simultaneous high ion mobility (conductivity) can lead to higher currents and consequently to Joule heating in the capillary, thereby reducing the separation resolution. This can be compensated for by lower capillary diameters or by use of zwitterionic buffers that can be adapted to the samples and used at high concentrations.

Temperature Temperature gradients across the capillary lead to loss in separation efficiency, but can be limited with an active cooling. Under extreme Joule heating conditions the liquid at the center of the capillary can approach the boiling point and, thus, can affect thermally labile analytes. Controlling the temperature thus enables reproducibility of separation conditions and, indirectly, the stability of the pK_a values, mobility, and solubility reaction equilibria. Lower buffer viscosity, at higher temperature, leads to faster separations. Reducing temperature enables the "freezing" of an ongoing interaction between buffer constituents and analytes; this is especially useful in the separation of enantiomers and isomers that may not be stable at higher temperatures. In any case a constant temperature over time is the rule for best reproducibility.

Capillary Coating The dissociation of silanol groups at the capillary surface is pH-dependent, enabling the EOF but also possibly inducing interactions with some analytes. This leads to peak broadening and loss of resolution and reproducibility. Coating of the capillary surface may reduce these effects. The various modes of capillary surface modification include:

- chemical modification of the silanol groups,
- dynamic coating of the capillary wall with polymers,
- addition of zwitterionic detergents to the buffer system,
- increasing the ionic strength.

These approaches have had various success rates in hindering the analyte–wall interactions and may find limited application, especially with compounds of limited pH-stability such as proteins. These issues will be discussed below.

Dynamic coating is realized by rinsing the column with polymer or buffer-electrolyte solutions. These additives coat the capillary surface and alter the surface charge, thus reducing the EOF. Dynamic coating can be used with fused silica as well with chemical modified capillaries. Figure 12.10 shows such an example in the separation of five basic proteins in a capillary coated dynamically with a solution of 0.02% poly(ethylene glycol) (PEG). The high efficiency of the separation is reached because the interaction of the positively charged basic proteins with the capillary wall is reduced; the number of theoretical plates obtained in this way is in the range of millions.

Buffer Additives The addition of cationic detergents that adsorb by interaction with the negatively charged capillary wall induces an inversion of the EOF. Positively charged sample ions can no longer interfere with the capillary wall and can be separated more effectively. The increase in ionic strength (on increasing the buffer concentration) also reduces the interaction of the analytes with the capillary wall.

Furthermore, buffer additives can also be used to affect secondary equilibria involving the sample ions. *cis*-Diol groups such as, for example, those present in carbohydrates or polyphenols can interact with borate ions to form borate-ester complexes that bear a negative charge, which affects their UV absorbance. The sugar–boron complexes can thus be separated with capillary electrophoresis and detected in the UV-Vis detection mode. The separation can be optimized by finding the right pH and borate concentration (around the pK_a of tetraborate, which is 9.4).

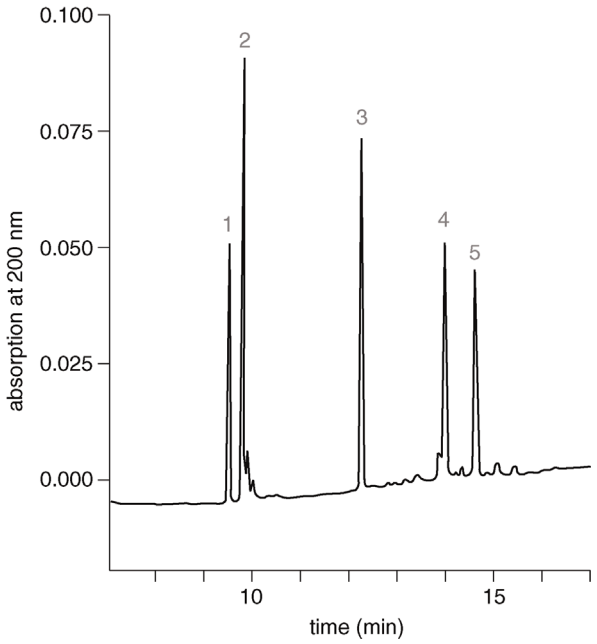
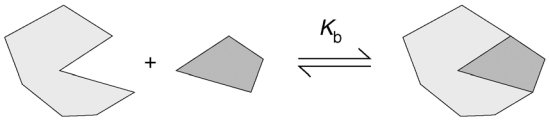


Figure 12.10 Separation of five alkaline proteins on a capillary dynamically combined with poly(ethylene glycol) (PEG). Buffer: 0.05 M β-alanine/acetic acid, incl. 0.02% PEG, pH 4.0 (1 cytochrome c, 2 lysozyme, 3 ribonuclease A, 4 trypsinogen, 5 chymotrypsinogen A).



mass	M	m	$M + m$
net charge	Z	z	$Z + z$
mobility	$\propto Z/M^{2/3}$	$\propto z/m^{2/3}$	$\propto (Z + z)/(M + m)^{2/3}$

Figure 12.11 Changing the mobility of a protein by formation of a chelate complex.

Very similarly, the separation of metal ions can be fine-tuned with the addition of chelators such as, for example, citric acid, lactic acid, or crown ethers.

The addition of various ligands, cyclodextrins or crown-ethers for chiral separations, and various micelle-building detergents or polymers sieves will be described below in the sections related to ACE, chiral separation, and MEKC or CGE, respectively.

12.4.2 Affinity Capillary Electrophoresis (ACE)

Affinity capillary electrophoresis (ACE) is a special form of CZE, in which interactions between proteins and ligands are used. Receptor–ligand interactions as well as binding constants and stoichiometry are important. Ligands can attach proteins and therefore form complexes, which have a different mobility than that of the protein itself provided that either the ligand carries a charge or the molecular weight of the complex is significantly different from that of the protein (Figure 12.11). Charge differences are mostly responsible for mobility alteration when using small ligands.

Typically, the mobility shift is determined by the addition of different ligand concentrations to the buffer solution at constant protein concentration (titration).

Complexation of monovalent protein–ligand complexes can be described by:

$$K_b = \frac{[P \cdot L]}{[P][L]} \tag{12.3}$$

where: K_b is the binding constant,
[P·L] is concentration of the complex,
[P] is the concentration of the protein
[L] is the ligand concentration.

The alteration of migration time for a definite ligand concentration is given by:

$$\delta t = t[L] - t[L]_0 \tag{12.4}$$

where t[L] is the migration time at a particular ligand concentration and $t[L]_0$ is the migration time without ligand addition.

The amount of protein, α, that is present as a complex is:

$$\alpha = \frac{\delta t_{[L]}}{\delta t_{max}} \tag{12.5}$$

where δt_{max} represents the maximum migration change, which means that saturation is reached. By introducing the binding constant K_b the following equation results:

$$\alpha = \frac{K_b[L]}{(1 + K_b)[L]} \tag{12.6}$$

The binding constant is determined by Scatchard analysis according to the following equation, which results from the above-mentioned expressions:

$$\frac{\alpha}{[L]} = K_b - \alpha K_b \quad \text{or} \quad \frac{\delta t}{\delta t_{max}} \frac{1}{[L]} = K_b \left(1 - \frac{\delta t}{\delta t_{max}}\right) \tag{12.7}$$

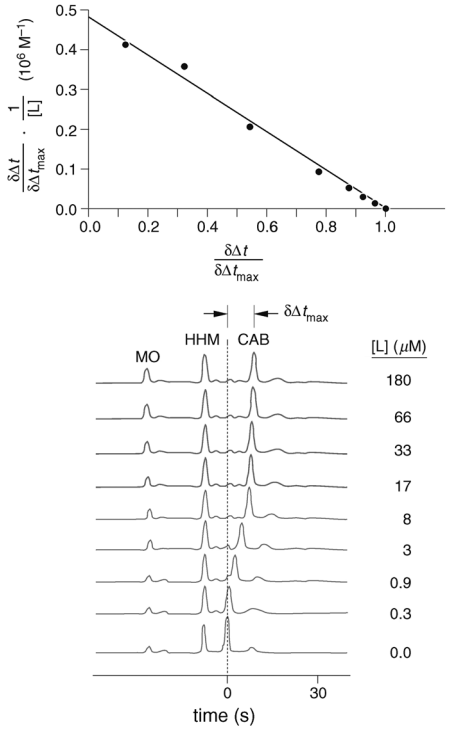


Figure 12.12 Affinity electrophoresis. Changing the mobility $\delta\Delta$t of carbonic anhydrase B (CAB) with variable concentrations of the ligand L (sulfonamide) and the corresponding Scatchard plot. Mesityl oxide (MO) and myoglobin (HHM) were used as internal standards. Affinity constant results are from the slope of the line or the *x*-intercept. Source: Chu, Y.H. *et al*. (1992) *J. Med. Chem.*, **35**, 2915–2917. With permission, Copyright © 1992, American Chemical Society.

As an example, the determination of the binding constant of alkyl benzene sulfonamide at carbonic anhydrase B is shown in Figure 12.12. This figure shows a series of electropherograms of carbonic anhydrase B at different concentrations of sulfonamide. The shift in the migration times in comparison to the migration time without addition delivers δt (in the figure $\delta\Delta t$, as the times are relative to the reference protein myoglobin in order to increase the precision of the measurement). The Scatchard plot $((\delta t/\delta t_{max})[L]^{-1})$ gives the binding constant, directly from the slope of the line or the intersection with the abscissa. This procedure requires that equilibrium occurs during the CE run and that the protein concentration is low enough to achieve saturation when using higher ligand concentrations. The binding constants determined in this way agree very well with those obtained by other methods.

12.4.3 Micellar Electrokinetic Chromatography (MEKC)

Micellar electrokinetic chromatography MEKC is a hybrid technology of electrophoresis and chromatography, developed in the early 1980s by S. Terabe. The addition of micelle forming agents (detergents) to the buffer system leads to the development of a pseudo-stationary phase, which consists of charged micelles. The separation of analytes is based on their distribution between the inside of the micelles and the solution (Figure 12.13). In principle, MEKC is therefore comparable with ACE. Neutral molecules, which are not separable by CZE as they do not migrate electrophoretically and are only transported by the EOF, obtain an electrophoretic mobility by their interaction with the charged micelles in MEKC. This electrophoretic mobility μ depends on the mobility of the micelles μ_{MC} and the capacity factor $k'i$:

$$\mu_i = \mu_{MC}\left(\frac{k'_i}{1+k'}\right) \tag{12.8}$$

The capacity factor k is obtained from the ratio of the residence time of the analyte in the mobile phase to that in the pseudo-stationary phase and can be simply calculated from the migration times of analyte and the micelle EOF (t_i, t_{MC}, and t_0):

$$k'_i = \frac{t_i - t_0}{t_0\left(1 - \frac{t_i}{t_{MC}}\right)} \tag{12.9}$$

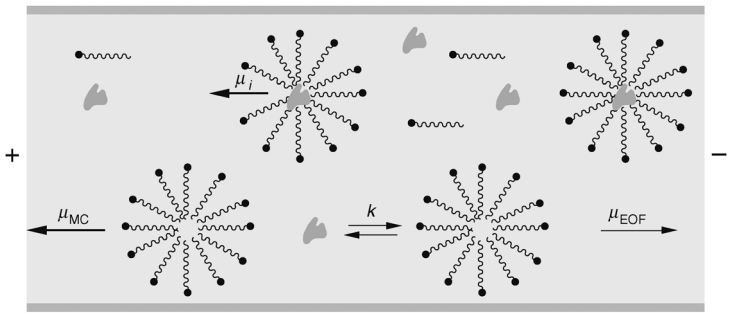


Figure 12.13 Principle of MEKC: Distribution of neutral analytes between background electrolyte and interior of the micelles. Mobility of uncharged ions in the electrolyte matches the EOF (μ_{EOF}) and inside micelles it is defined as μ_{MC}. This results in an effective mobility μ_i, which depends on distribution coefficient k.

where t_0 corresponds to the migration time of an unretained component and can be determined by the "EOF peak," which is visualized by the refractive index change of the buffer; t_{MC} is determined by the usage of hydrophobic substances (like Sudan III). Such substances have an extremely high k values and are present only inside the micelles.

The analytes dependency on polarity/charge causes them to show a different affinity to the pseudo-stationary phase of micelles. They differ therefore in their average residence times inside those micelles and, as a consequence, in their migration times. The migration times of all analytes show values between the EOF and migration times of the micelles themselves (Figure 12.14). An example is given in Figure 12.15. Here, dansyl-derivatized amino acids are separated in an abundance of SDS as the micelle building agent. Choices of micelle building agent and modification in the composition of the aqueous phase have the biggest impact on

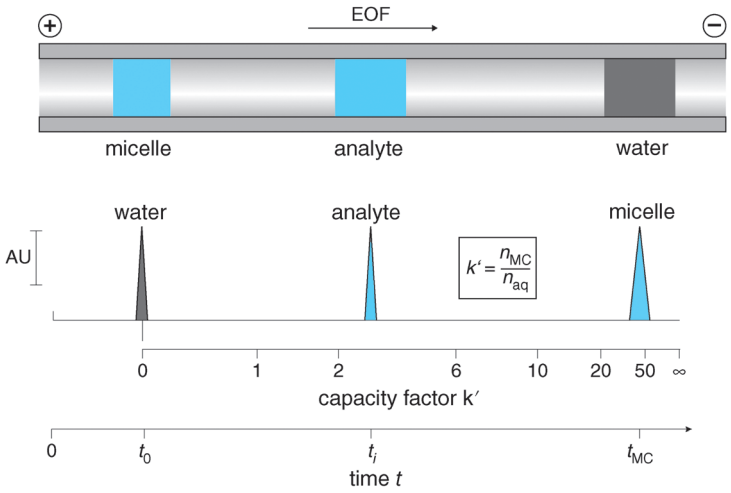


Figure 12.14 Time window of MEKC. The migration time depends on capacity factor k' and is limited to the range between EOF-t_0 and t_{MC} (migration time of micelle).

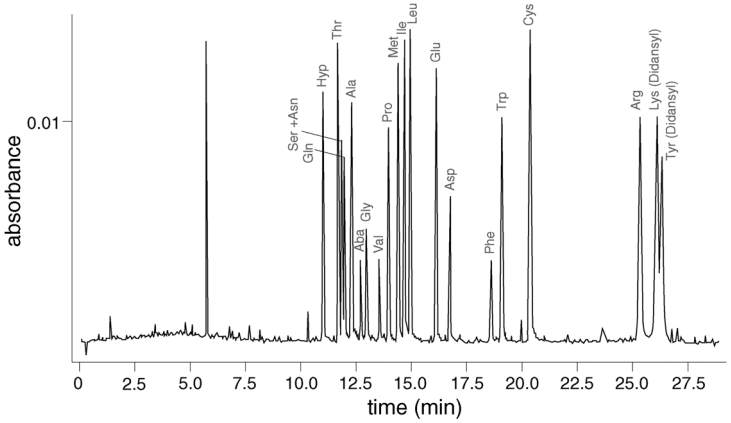


Figure 12.15 Separation of dansyl derivatives of amino acids in the presence of SDS to form micelles above the CMC. Source: Miyashita, Y. and Terabe, S. *Application Data, High-Performance Capillary Electrophoresis, Beckman, Fullerton, CA,* (1990) Beckman DS-767.

Table 12.3 Critical micelle concentration (CMC) and aggregation number n in water at 25 °C.

Micelle former	CMC (in 10^{-3} M)	*n*
Anionic:		
sodium dodecyl sulfate (SDS)	8.1	62
sodium tetradecyl sulfate (STS)	2.2	138
sodium cholate (salt of bile acid)	13–15	2–4
sodium taurocholate	10–15	5
Cationic:		
cetyl(trimethyl)ammonium bromide (CTAB)	0.92	61
dodecyl trimethyl ammonium bromide (DTAB)	15	56
Zwitterionic:		
3-[(3-cholamidopropyl)dimethylammonio]-1-propane sulfonate (CHAPS)	4.2–6.3	9–10

resolution due to their modification of selectivity. The addition of organic solvents causes a decrease of capacity factor and a decrease of EOF.

Micelle building agents might be anionic, cationic, or zwitterionic compounds. All are bipolar in nature with a hydrophilic and a hydrophobic molecular part. If their concentration is high enough, micelle molecules will attach to each other with the hydrophobic tails pointing towards the inside and the hydrophilic heads facing the aqueous solution. The critical concentration, termed the critical micelle concentration (CMC), is a detergent-specific parameter as well as the aggregation number. This is the number of molecules in each micelle. A typical micelle size is 3–6 nm; consequently, it is a homogeneous solution. Important attributes for micelle building agents are:

- high solubility in buffer (≫CMC)
- low UV-absorption
- low viscosity.

Detergents, Section 1.8

Table 12.3 shows a selection of different detergents and their critical micelle concentration as well as their aggregation number. SDS is the most often used micelle builder. The example given in Figure 12.15 shows the MEKC separation of dansylated amino acids in SDS-buffer.

Notably, quaternary ammonium salts with C_{10} to C_{18} long alkyl chains absorb even below their CMC strongly onto the fused silica capillary wall, which causes a flow reversal. The negative surface causes an EOF, which is directed towards the anode. Anions will thus follow this EOF and their electrophoretic migration will cause them to reach the detector before neutral molecules and cations (polarity reversal!).

12.4.4 Capillary Electrochromatography (CEC)

Capillary electrochromatographic (CEC) combines the advantages of HPLC and CE in one separation method: the phase separation of chromatography and the electrokinetic principles of electrophoresis. HPLC is the most widely used separation technique in bioanalysis and includes a large variety of stationary phases ranging from polar to very nonpolar materials (Figure 12.16). These materials are placed into capillaries and are used as stationary phases in electrochromatography. However, a big problem is the formation of gas bubbles during the separation, caused by the application of a voltage.

Chromatographic Separation Techniques, Chapter 10

Because of the ability to control the surface chemistry and porosity, nowadays monolithic polymer columns have prevailed over particle-packed and sol–gel capillaries for CEC.

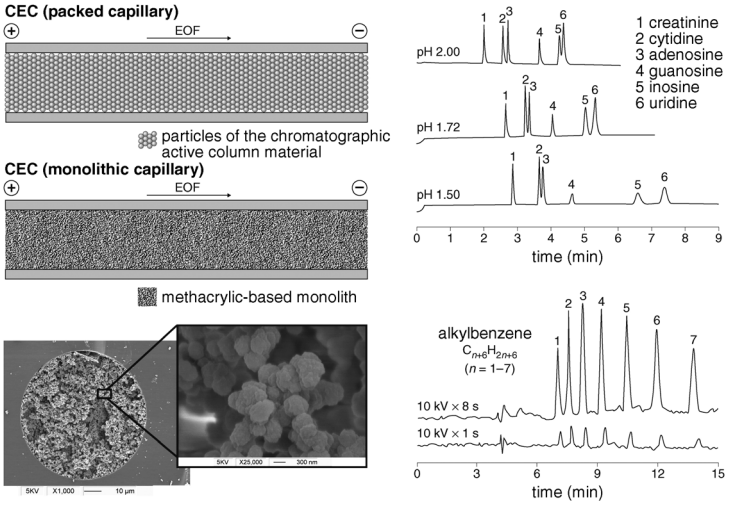


Figure 12.16 Example of capillary electrochromatographic (CEC) separation of polar and nonpolar substances in methacrylic-based monolithic capillaries. Source: Ping, G., Zhang, W., Zhang, L., Zhang, L., Schmitt-Kopplin, Ph., Kettrup, A., and Zhang, Y. (2003) *Chromatographia*, **57**, 777–781 With permission, Copyright © 2003, Friedr. Vieweg & Sohn Verlagsgesellschaft mbH.; Ping, G., Zhang, Y., Zhang, W., Zhang, L., Zhang, L., Schmitt-Kopplin, Ph., and Kettrup, A. (2004) *Electrophoresis*, **25**, 421–427. With permission Copyright © 2003 Wiley-VCH Gmbh & Co. KGaA.

12.4.5 Chiral Separations

In CE, because most chiral molecules are neutral, chiral separations are usually carried out using MEKC; however, chiral molecules that are natively charged can also be separated by CZE or CEC. In addition to the detergent, an enantioselective reagent, for example, cyclodextrins (CDs), can be added to the buffer system. Cyclodextrins are produced by the enzymatic degradation of starch. They consist of six to eight glucose units (α-, β-, γ-CD) arranged in the form of a molecular capsule. The hydrophilic OH groups are located on the outside, while the hydrophobic hydrocarbon skeletons are directed inwards (Figure 12.17). The D-glucose units give the CD molecules an inherent molecular asymmetry, so that chiral guest molecules can be recognized enantioselectively and attached.

The enantiomeric mixtures of chiral analytes are distributed between the inside of the selective CD and the interior of SDS micelles in the buffer if their spatial arrangement fits. The enantiomer interacting strongest with the SDS micelles moves slower than the cyclodextrin-captured enantiomers due to the negative charge of the SDS, which moves with higher speed in the opposite direction to the EOF (Figure 12.18).

Chiral MEKC is not used very often in bioanalysis. However, one example is the separation of derivatized D- and L-amino acids. A broad range of applications for chiral CE separations exists in the analysis of chiral pharmaceutical ingredients, as well as of chiral pesticides.

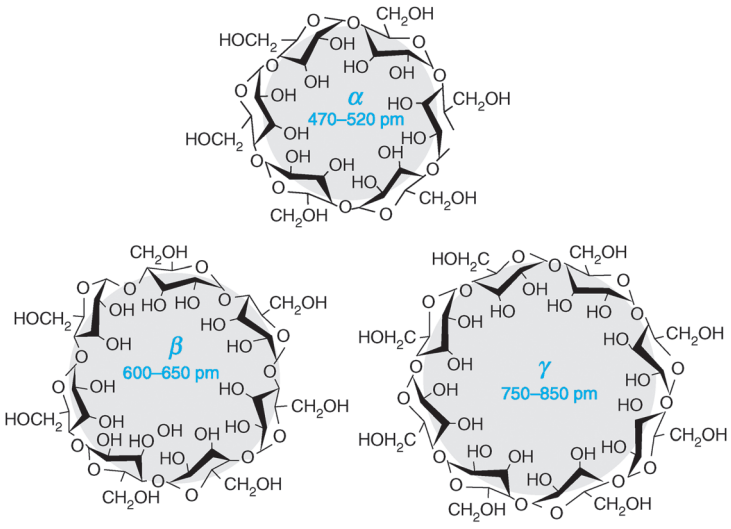


Figure 12.17 Structure and size of α-, β-, and γ-cyclodextrin.

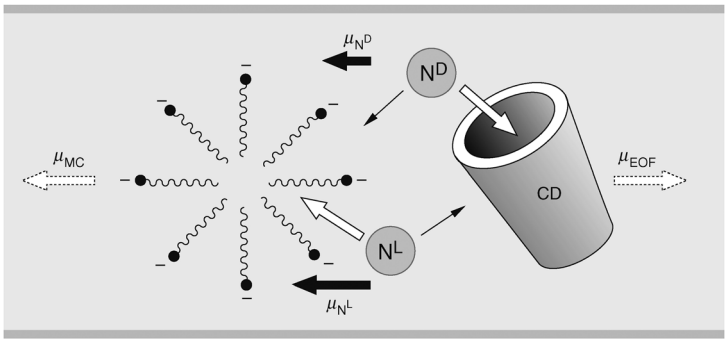


Figure 12.18 Principle of chiral MEKC. Weaker interactions of the enantiomer of an uncharged chiral analyte (N^L) with cyclodextrin results in increasing probability of its presence in the micelle and thus in more effective mobility. Uncharged cyclodextrin is only transported by EOF; therefore the mobility of chelating enantiomer is reduced.

12.4.6 Capillary Gel Electrophoresis (CGE)

The most frequently used electrophoresis technique is the slab-gel electrophoresis of proteins and DNA in biochemical and molecular biology laboratories. Classically, gels were used as anticonvective media in electrophoresis to avoid peak broadening due to temperature. Since very thin capillaries with efficient heat dissipation are now used in capillary electrophoresis, there is no need to use such media. However, gels can be also used as sieving media. In capillary electrophoresis, crosslinked or linear gels are used. CGE is a special form of CZE: in CZE separation takes place according to different mass/charge ratios, that is, ion mobilities. But, both DNA molecules and SDS-denatured proteins might have very similar mass/charge ratios even if their molecular masses are different, so they cannot be separated by CZE. This can be overcome by utilization of sieve effects. Separation according to size becomes possible and the gel medium hinders the electrophoretic migration of larger molecules more than those of smaller ones. Gel electrophoresis in capillaries has several advantages over classical slab-gel electrophoresis.

In addition to crosslinked polyacrylamide gels, linear polymers are also used as a sieving medium in CGE (Table 12.4).

Crosslinked gels are composed of two monomer units (acrylamide and bisacrylamide). They show a defined pore size and they are very rigid in their physical properties. These gels are polymerized in the capillaries, and they are covalently bound to the capillary wall (chemical gels). Often, gas bubble formation and shrinkage of the gel can be observed during polymerization. The gels are not interchangeable and have only a limited lifespan (about 100 separations), but their separation performance is exceptional.

Linear gels consist of a loose mesh of linear polymer chains that is held together only by physical interactions (physical gels). The highly viscous gel buffer solutions are interchangeable with the gel column contents at high pressure, so that after each separation a new gel can

Table 12.4 Sifting media for CGE.

Polymer	Concentration (%)	Application
Crosslinked polyacrylamide	2–6 T [a]	Oligonucleotides, DNA sequencing
	3–6 C [b]	
Linear polymers:		
Polyacrylamide	6–10	Oligonucleotides
	<6	restriction fragments, PCR fragments
Cellulose derivatives	<1	PCR fragments
Poly(ethylene glycol)	<3	Proteins
Dextran	10–15	Proteins
Agarose	<1	Proteins, restriction fragments

a) Total concentration of acrylamide.
b) Crosslinking level.

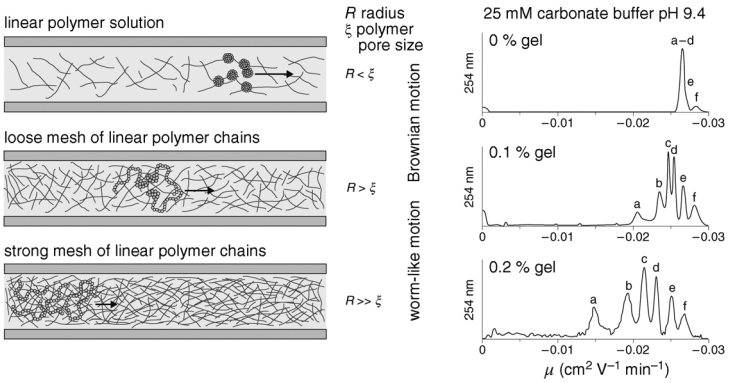


Figure 12.19 The concentration of linear polymer solution influences the sifting effect and CGE separation of polystyrene sulfonate polymers of various sizes. Source: Schmitt-Kopplin, Ph. and Junkers, J. (2007) *Environmental Colloids: Behavior, Structure and Characterization* (eds K. J. Wilkinson and J.R. Lead), IUPAC Series on Analytical and Physical Chemistry of Environmental Systems, vol. 10, John Wiley & Sons Ltd, Chichester, pp. 277–314. With permission.

theoretically be filled into the capillaries. However, only solutions of up to about 4% linear polyacrylamide are replaceable with commercial equipment. The distinction between so-called "fixed" and "dissolved" gels is therefore mostly semantic.

For the separation of proteins, the use of polyacrylamide as the sieving matrix is problematic due to the absorption of the polyacrylamide itself below 230 nm. However, the sensitivity is not decreased by using a wavelength of 280 nm compared to 214 or 200 nm. Therefore, dextrans or poly(ethylene glycol)s with molecular weights in the range of 100 000 amu are the UV-transparent sieving media of choice for routine applications (Figure 12.19).

The relationship between migration time and molecular mass of the SDS–protein complexes shows excellent linearity. The method is suitable for routine use in the analytical sector and can even replace SDS-PAGE. The molecular weight window ranges from about 15 to about 200 kDa.

12.4.7 Capillary Isoelectric Focusing (CIEF)

Classic capillary isoelectric focusing CIEF is nowadays indispensable in protein analysis. The main disadvantages of gel IEF are the lack of opportunity for automation, long time consumption, and poor quantification possibilities due to the staining reaction for detection. Variations during gel production often lead to problems in reproducibility. The transfer of isoelectric focusing to the capillary can solve these problems, but requires an adaptation of the instrumental conditions.

Isoelectric Focusing, Section 11.3.12

As in commercial CE devices detection occurs only at one fixed point; proteins must be mobilized after their focusing, meaning that they need to be transported through the detector. There are three methods of mobilization:

- one-step focusing with mobilization by the EOF,
- focusing with pressure/voltage mobilization (Figure 12.20),
- focusing with chemical mobilization.

The pH gradient is formed by a large number of ampholytes with different pI values. The smaller the spacing between the individual pI values, the more homogeneous is the pH gradient. Typically, the entire capillary is filled with a mixture of ampholytes and sample. When a voltage is applied the ampholyte ions start to migrate according to their pI values and establish as a consequence a pH gradient. When a pH value corresponding to the pI value of an analyte is reached, its electrophoretic migration ends and so a decrease of electric current is measured. In this pH gradient a protein will migrate until its velocity becomes zero, which means that it has reached a pH that is equal to its pI value.

Since the on-column detector is always a certain distance from the capillary and mobilization occurs only in one direction, the pH gradient should be formed just before the detector in order not to lose the strongly acidic or strongly basic proteins. This can be achieved by blocking the other end (seen from the detector) of the capillary, either through the catholyte (NaOH) or by the

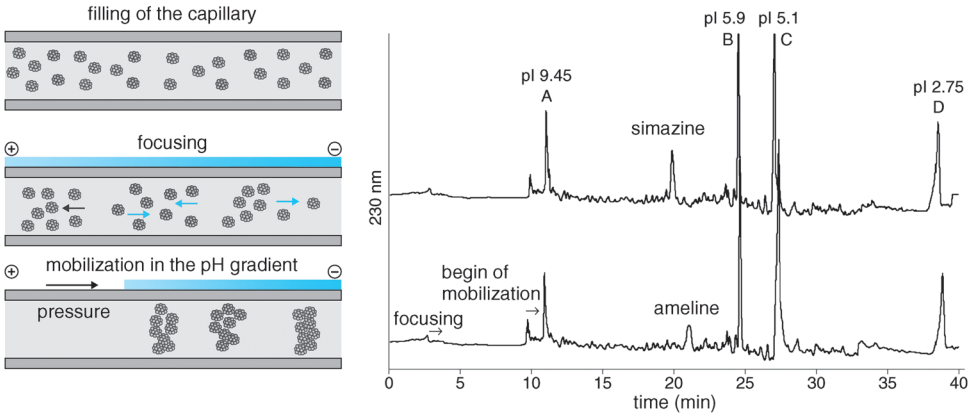


Figure 12.20 Principle of isoelectric focusing with pressure mobilization. The capillary is completely filled with a mixture of ampholyte and protein. By applying a voltage to the capillary a pH gradient between NaOH and H_3PO_4 (on opposite sides of capillary) is formed and the focusing of proteins is effected in parallel. Afterwards analyte proteins are mobilized to be detected. A–D are CIEF separated proteins with different pI values. Source: Schmitt, Ph., Poiger, T., Simon, R., Garrison, A.W., Freitag, D., and Kettrup, A. (1997) *Anal. Chem.*, **69**, 2559–2566. With permission Copyright © 1997 American Chemical Society.

addition of *N*,*N*,*N*′,*N*′-tetramethylethylenediamine (TEMED) to the ampholyte mixture. TEMED as a very basic compound that migrates towards the basic end of the pH gradient, blocking one portion of the capillary for the pH gradient. The part of the capillary from the detector to the capillary end can be occupied by TEMED very accurately by choosing the TEMED concentration. This solution is experimentally very easy and is reproducible. Other methods make use of the lower diffusion in viscous media or by use of gels, as described in Section 13.4.6.

One-Step Focusing Simultaneously with the focusing process, mobilization occurs, as the EOF is only reduced but not completely eliminated. Since the pH-dependency of EOF results in a decrease of the EOF during the mobilization, the linearity of the pI calibration is lost. This EOF decrease during the mobilization can be reduced when the pH gradient is restricted to only a short length of the capillary (the capillary end to the detector) and the other longer part of the capillary is blocked with TEMED. As a consequence, the pH in the main part of the capillary remains basic during the mobilization.

The advantages of this method are ease in handling and short analysis time. In particular, the method is well suited for the pH range of about 8.5 to 4.5 due to its robustness and speed, for example, for the separation of monoclonal antibodies (Figure 12.21). However, for very basic proteins the method is less suitable, since these are often not fully focused when they pass through the detector.

Focusing with Pressure/Voltage Mobilization Focusing and mobilization can be separated when the EOF is eliminated. Proteins are fully focused in a first step and are thereafter pushed through detector by the application of pressure. As this occurs at a constant speed, linearity of the calibration curve is maintained. Figure 12.22 shows the focusing of standard proteins covering a wide range of pI values.

To maintain the high resolution of focusing, it is absolutely necessary to apply high field strength at the capillary during mobilization so that no mixing caused by the hydrodynamic flow profile occurs. The resolution obtained by pressure mobilization is only at high field strengths relative to those used for one-step focusing.

Focusing with Chemical Mobilization Focusing and mobilization are realized separately, which requires a complete suppression of the EOF. The mobilization is achieved chemically by changing the composition of the analyte or catholyte. For instance, in cathodic mobilization the OH^- ions are exchanged with Cl^- ions, resulting in a reduction of the pH. Following this destruction of the pH gradient proteins are no longer located at a pH corresponding to their pI values. Thus, they become positively charged and begin to migrate electrophoretically toward the cathode. As a consequence, they can be detected. The linearity of

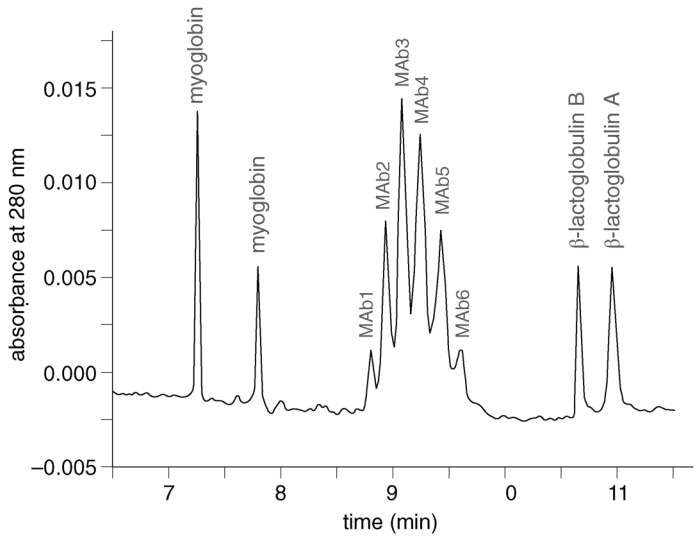


Figure 12.21 One-step focusing of a monoclonal antibody having internal pI markers. Source: Schwer, C. (1995) *Electrophoresis*, **16**, 2121–2126. With permission Copyright © 1995 VCH Verlagsgesellschaft mbH.

this method is very good; only the last mobilized, acidic part is slightly compressed. Therefore, for acidic proteins anodic mobilization is advantageous. This can be done, for example, with Na^+ ions.

Analysis times can be shortened by increasing the concentration of the mobilizing ions. But this leads – to a limited extent – to a decrease of resolution, because higher concentrations cause an increase in Joule heat generation.

Focusing with chemical mobilization provides the highest resolution of the three mobilization methods, because neither a hydrodynamic flow nor an uneven EOF causes a loss of efficiency. The example with a very basic monoclonal antibody illustrates this (Figure 12.23).

12.4.8 Isotachophoresis (ITP)

Separation in isotachophoresis (ITP) is based on mobility, that is, the size and charge of the ions, analogous to that in CZE. The difference lies in the electrolyte assembly. While in CZE the entire capillary is filled with a carrier electrolyte, causing a constant field strength, in ITP an arrangement of two electrolytes is used: one supporting electrolyte and one terminating electrolyte. The mobility of the supporting electrolyte must therefore be higher than the mobility of all analytes and the mobility of the terminating electrolyte must be lower.

Following Ohm's law, a constant current leads to the formation of a field strength gradient. This is due to the different mobilities and therefore conductivities of the two electrolytes. In the

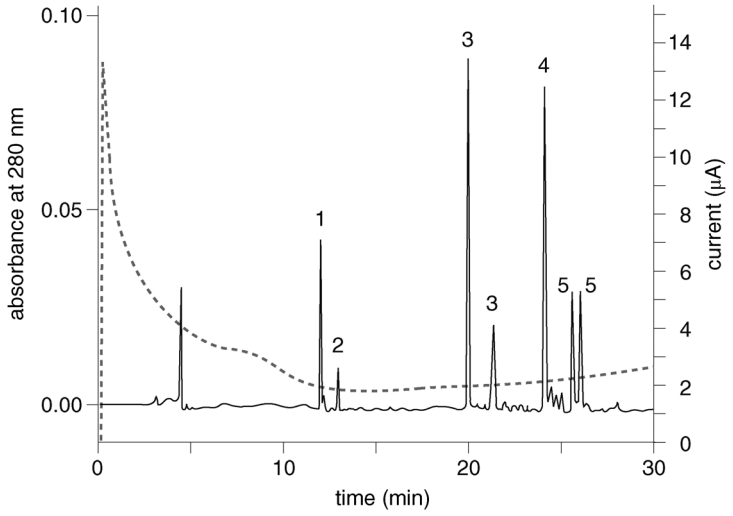


Figure 12.22 CIEF with pressure/voltage mobilization of standard proteins. 1 cytochrome *c*, 2 ribonuclease A, 3 myoglobin, 4 carbonic anhydrase, 5 β-lactoglobulin A, B. Source: Schwer, C. (1995) *Electrophoresis*, **16**, 2121–2126. With permission Copyright © 1995 VCH Verlagsgesellschaft mbH.

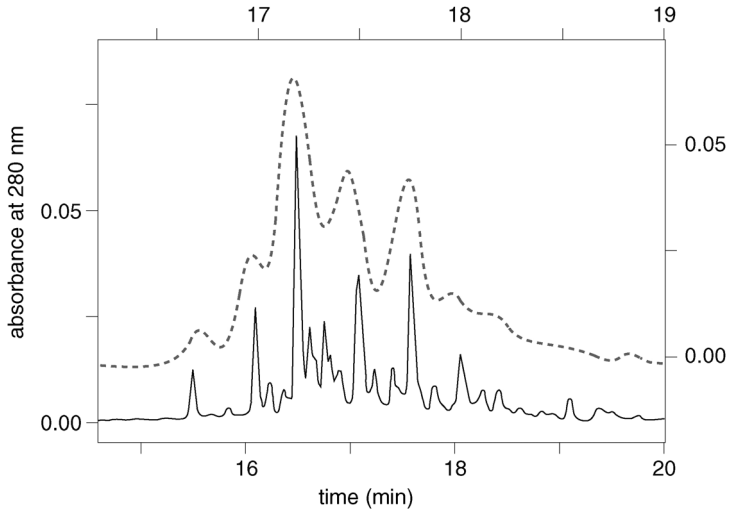


Figure 12.23 Comparison of CIEF of a basic monoclonal antibody using pressure (red dashed line) or chemical mobilization (solid line). Source: Schwer, C. (1995) *Electrophoresis*, **16**, 2121–2126. With permission, Copyright © 1995 VCH Verlagsgesellschaft mbH.

zone of the supporting electrolyte, there is a lower field strength than in the zone of terminating electrolyte. The sample is injected at the interface of the two electrolytes. According to the mean conductivity of the zone there is an averaged field strength E_{mix}. In this mixed zone the analyte ions migrate with different speeds depending on their mobility:

$$v_i = \mu_i E_{mix} \tag{12.10}$$

Fast ions accumulate at the front and slower ones at the end of the zone. A steady state is reached. Each analyte ion forms thereby a separate zone, which is characterized by a field strength corresponding to the mobility (Figure 12.24). All zones migrate now at constant speed, which is given by the speed of the supporting electrolyte.

The zones are not broadened by diffusion as in CZE. In contrast they remain sharp due to the "self-sharpening" effect caused by the field strength gradient. Different field strengths in other zones result in either the slowing down or acceleration of diffusing ions. The concentration profile is rectangular, while the relation of concentration to the supporting electrolyte follows

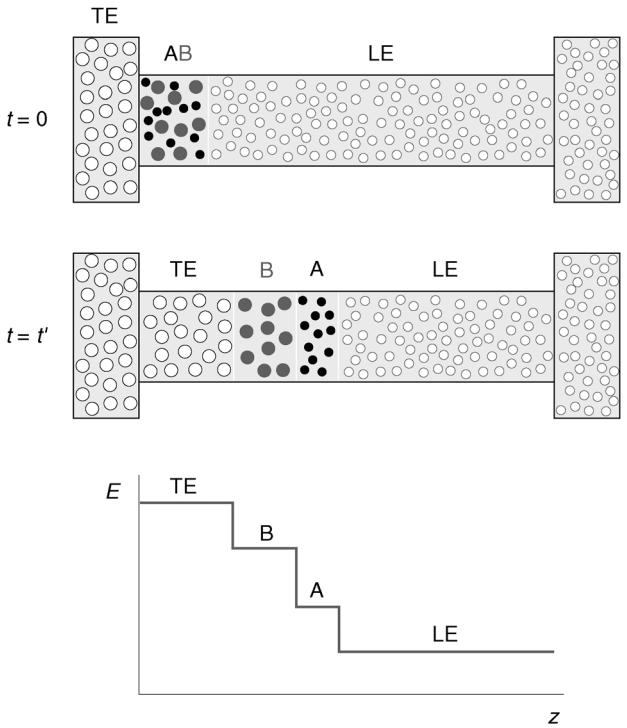


Figure 12.24 Order of electrolytes and course of field intensity within isotachophoresis (ITP). The system consists of a leading electrolyte (LE) and a terminating electrolyte (TE), whereby ionized analytes (A, B) are introduced between both zones. At *t'* ions A and B are separated and form their own zones with field intensities corresponding to their mobility. This stepwise gradient of field intensities prevents diffusion of zonal borders.

the Kohlrausch equation:

$$C_A = C_L \frac{\mu_A(\mu_L + \mu_Q)}{\mu_L(\mu_A + \mu_Q)} \tag{12.11}$$

where c_A and c_L are concentrations of the analyte and the supporting electrolyte, respectively, and μ_A, μ_L, and μ_Q are the mobilities of analyte, supporting electrolyte, and counter-ion, respectively.

The analyte zones incorporated no foreign ions except the counter-ions. In contrast to CZE, in which dilution of the sample zone occurs by diffusion, diluted samples are enriched in ITP.

12.5 Special Techniques

12.5.1 Sample Concentration

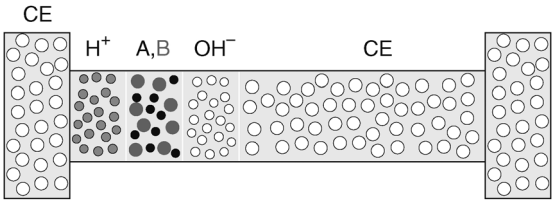
Although not part of the CE instrumental analysis phase, one should not forget the simple concepts of sample pretreatment for analyte enrichment. Usually, real-world samples are too low in concentration for direct analysis by CE, at least for routine CZE analysis with UV-Vis detectors. However, the analytes of interest have usually been extracted from the natural matrix by an organic solvent and this solvent can readily be reduced carefully in volume by simple evaporation with minimal loss of analyte. The solvent cannot usually be injected directly into the CE, but must be exchanged for water. This can be accomplished by dissolution of the dried extract residue in water, or for highly hydrophobic analytes by dissolution in a small amount (e.g., 50–100 µl) of methanol or acetonitrile, which can be made to the final known volume with about a tenfold volume of distilled water. The CE technique can usually accommodate up to 25% of methanol in the injected aqueous sample solution; this does affect the migration time of analytes somewhat, so analyte standards solutions have to be of the same composition.

12.5.2 Online Sample Concentration

Capillary electrophoresis requires and allows only very small sample volumes. Only tiny sample volumes can be injected to retain the high separating efficiency of the capillary electrophoresis. Nevertheless, the required concentration for UV detection is relatively high (about 0.1 mg ml^{-1}). It is possible to increase the injection volume by 50- to 100-fold with stacking to enhance the concentration sensitivity, comparable to procedures used in classical SDS-PAGE. For this purpose, several discontinuous buffer systems have been developed.

One Buffer Stacking System This concentration system is particularly suitable for the concentration of amphoteric compounds. The sample zone is between two zones with extreme pH values (diluted NaOH and H_3PO_4) and is thereby "focused" (Figure 12.25). The subsequent separation is carried out, for example, in phosphate buffer.

Two Buffer Stacking System In this technique, the electrolyte system consists of two buffers, which are selected so that the separation buffer has a very low mobility (terminating



CE: 0.02 M Na-phosphate, pH 2.8
H^+: 0.1 M H_3PO_4
OH^-: 0.1 M NaOH
A, B: sample ions

Figure 12.25 Order of electrolytes within a single buffer stacking system. Ionized analytes are focused between H^+ and OH^-.

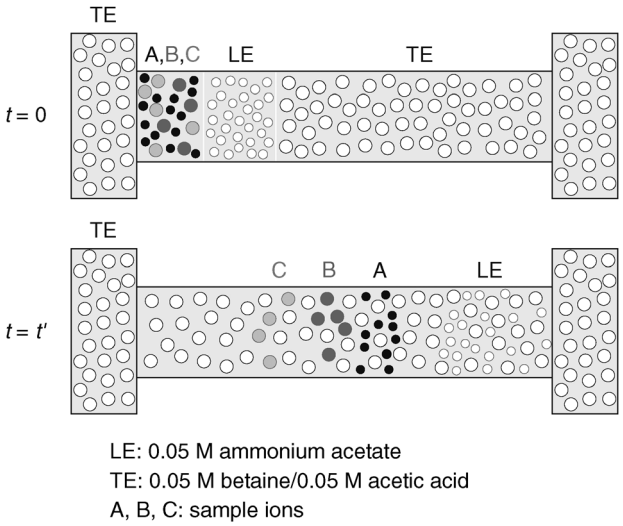


Figure 12.26 Order of electrolytes within a double buffer stacking system. The leading electrolyte (LE) in front of the analyte zone (see isotachophoresis) concentrates ionized analytes. At t' ions of the LE are migrating to the terminating electrolyte (TE), indicating that ionic interactions at zonal border are decreasing, followed by separation of analytes by zone electrophoretic conditions.

electrolyte) and the supporting electrolyte has a very high mobility. The capillary is filled with terminating electrolyte. In front of the sample, the supporting electrolyte is brought into the capillary. The capillary ends are dipped into the terminating electrolyte. At the beginning of the separation isotachophoretic conditions are present. The sample is concentrated between the supporting and terminating electrolytes. The ions of the supporting electrolyte migrate zone-electrophoretically away from the sample ions, and are eventually also separated (Figure 12.26). Ions of high mobility, which are in the sample solution, cause a similar effect to that of the supporting electrolyte zone in front of the sample.

12.5.3 Fractionation

Although several commercial devices offer the possibility to fractionate the separated analytes for further characterization, some crucial aspects need to be considered. Since the detector is not located at the end of the capillary and different analytes do not move with constant speed after the peak detection towards the capillary end – as is the case in HPLC – for each individual peak the time window for the fractionation has to be calculated. Field gradients in the system (e.g., caused by concentration steps carried out to increase the sample loading) lead in a non-constant migration rate and thus complicate the accurate time prediction if only the detection time is known. For very complex separations (e.g., tryptic cleavage of a protein) an arrangement that enables an automatic fractionation without prior determination of migration speeds is therefore desirable.

By supplying a make-up flow of 5–10 µl min^{-1} via a T-piece upstream of the detector the transport of the analyte after the detection is determined by this make up flow and is therefore constant. Samples can be collected continuously, without interrupting the power supply.

A sample concentration that is sufficient for subsequent sequence analysis can be reached by isotachophoresis. Figure 12.27 shows an electropherogram of a micropreparative separation of tryptic peptides. Eleven peaks were collected; the purity of each peak was evaluated by reinjection of a small portion (inset in Figure 12.27). All baseline-separated peaks showed a purity of 95%. The amounts of peptide from one separation were sufficient for subsequent characterization by amino acid sequence analysis and also for accurate molecular mass determination via laser desorption MS (MALDI-MS). In contrast to ESI-MS, MALDI-MS has the advantage that the capillary electrophoresis running buffer does not interfere with the analysis and shows a greater mass sensitivity. However, the main disadvantage is that it cannot be coupled online, thus requiring a fractionation.

MALDI-MS, Section 15.1.1

The required MALDI matrix solution can be fed in via a T-piece. This can be done directly during application of the fractions onto the MALDI plate. Furthermore, the generated spot is available for repeated analysis, which allows for MS/MS experiments (MALDI-TOF/TOF) and the possibility to fragment and identify more precursor ions at a later stage. In contrast, this has to be done simultaneously with the separation when using online ESI systems.

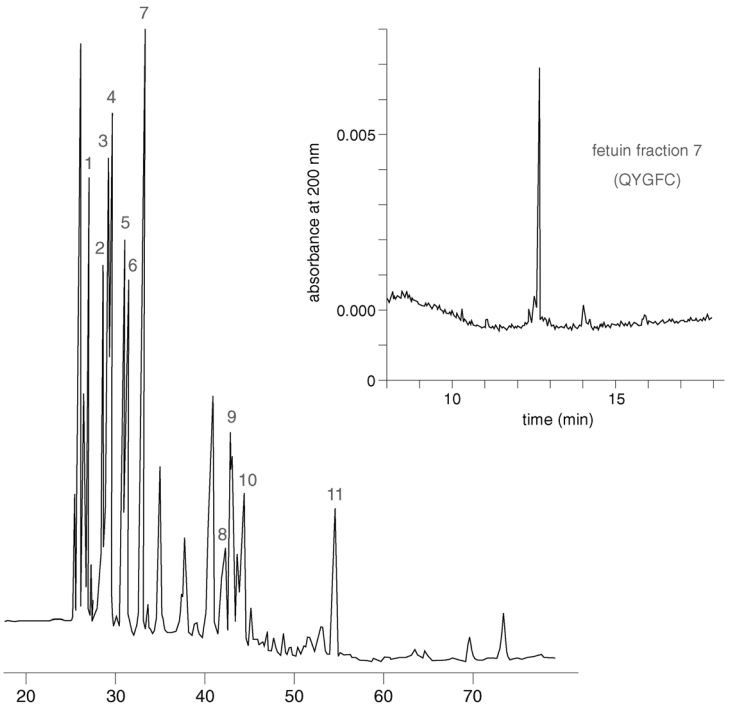


Figure 12.27 Micropreparative separation of tryptic fetuin peptides within betaine/acetic acid buffer (pH 3.3). Fractions of numerical peaks were collected and the purity was proved by reinjection. Insert: purity control of fraction 7. Isotachophoretic online sample concentration was performed to increase sensitivity. Amino acids 1–5 (QYGFC) were determined by sequence analyses.

12.5.4 Microchip Electrophoresis

Further miniaturization of capillary electrophoresis has led to the development of microchip electrophoresis (MCE). The motivation for this was – among other things – the reduced dimensions of the device and the improved performance capability of miniaturized systems in terms of high separation efficiency, short analysis times, and low consumption of solvents, as well as low sample volumes. The very low analysis times of MCE enable high-throughput screening or process analysis with maximum separation. One record – among many – is the separation of a binary mixture in 0.8 ms at 53 kV cm^{-1} in a capillary of 200 µm i.d. Micro-total-analysis systems (µ-TAS) as on-chip analytical laboratories of the future illustrate the possibilities and the analytical potential of MCE. All work steps such as sample preparation, derivatization, separation, and detection can be incorporated in one microchip.

Microchannels, in which the electrophoretic separations take place, are produced with photolithographic techniques. The separation channel is filled with the buffer (electrokinetically). The sample is filled electrokinetically during a second step (Figure 12.28); the amount of sample is thereby limited by the channel sizes (pinched injection). In contrast to this pinched injection, variable volumes can be inserted with the gated injection method (variations in injection times).

The surface of separation channels can be coated with different materials as in CZE, CGE, or MEKC. Examples of amino acid separation with a MEKC-chip and protein separation with chip-CGE are given in Figure 12.29. The fast separations are particularly suitable for two-dimensional separating methods, such as MEKC × CZE for separation of a digested protein. The detection is carried out as described in Section 12.3.4. Of particular importance is the very sensitive detection methods such as laser induced fluorescence (LIF), mass spectrometry, or classical UV-VIS detection.

12.6 Outlook

Bioanalysis is certainly only a small part of the total applications of capillary electrophoresis, which allows the analysis of small ions, metals, peptides, and proteins but also entire particles such as microorganisms and individual cells. Although gel electrophoresis and chromatography

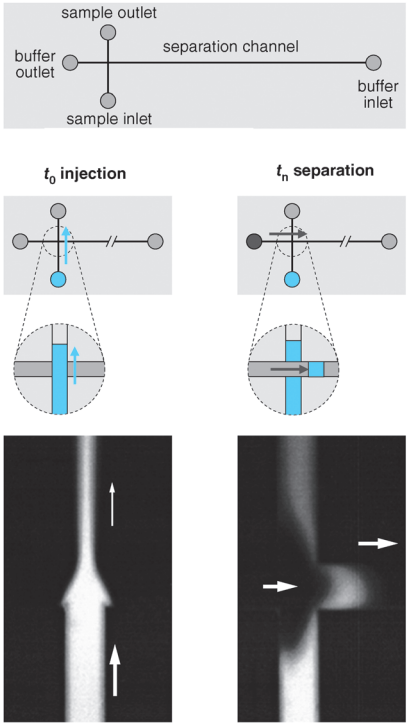


Figure 12.28 Miniaturization as microchip electrophoresis (MCE) – principle of injection and separation. Source: Ping, G., Zhang, Y., Zhang, W., Zhang, L., Zhang, L., Schmitt-Kopplin, Ph., and Kettrup, A. (2004) *Electrophoresis*, **25**, 421–427. With permission, Copyright © 2004 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

in conjunction with mass spectrometry are leading technologies in the field of protein analysis, great efforts are being made to employ other miniaturized techniques to avoid the disadvantages of 2D gel electrophoresis (low sample throughput, small dynamic range, non-detection of hydrophobic, highly acidic and basic, and very large proteins). For this purpose it is necessary to couple various techniques like capillary electrophoresis and chromatography and to use mass spectrometry for identification. Selective enrichment methods are further indispensable in order to detect very low levels of expressed proteins. There is a demand for robust procedures that can cope with the smallest sample quantities and allow routine analysis.

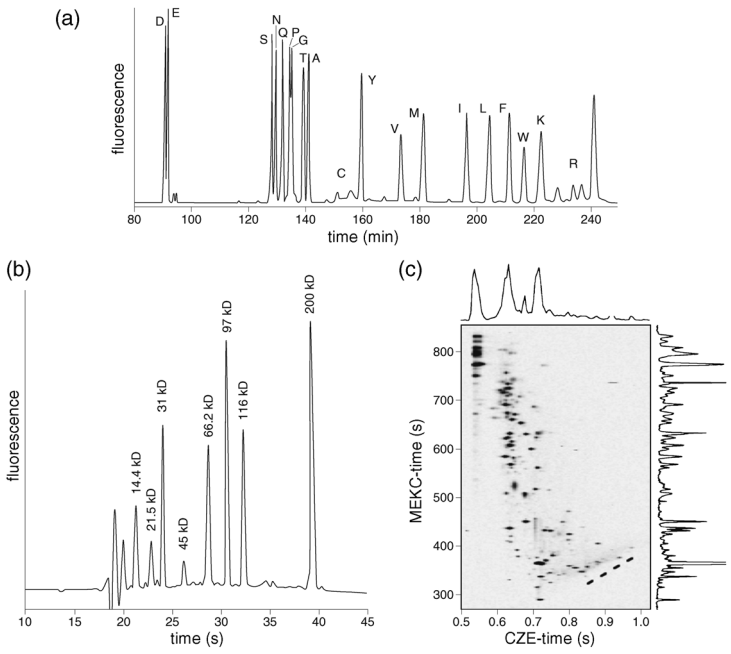


Figure 12.29 MCE separation of amino acid derivatives by chip-MEKC (a) and proteins by chip-CGE (b); multidimensional approach for separation of a tryptic protein (bovine serum albumin, BSA) by MEKC and CZE (c).

Other technical developments are the use of multiple parallel capillaries to increase sample throughput (an important advance in genome sequencing), further miniaturization (microchips), improved detection, and easier fraction collection possibilities.

Capillary electrophoresis can furthermore contribute to investigation of protein–ligand interactions or the structural behavior of proteins.

Further Reading

Ahuja, S. and Jimidar, M.I. (eds) (2008) *Capillary Electrophoresis Methods for Pharmaceutical Analysis*, Academic Press, Elsevier, London.

Altria, K.D. (ed.) (1996) *Capillary Electrophoresis Guidebook*, Humana Press, Totowa.

Bartle, K.D. and Myers, P. (eds) (2001) *Capillary Electrochromatography*, The Royal Society of Chemistry, Cambridge.

Camilleri, P. (ed.) (1998) *Capillary Electrophoresis*, CRC Press, Boca Raton.

de Jong, G. (ed.) (2016) *Capillary Electrophoresis - Mass Spectrometry (CE-MS) Principles and Applications 1* Wiley-VCH Verlag GmbH, Weinheim.

Deyl, Z. and Svec, F. (eds) (2001) *Capillary Electrochromatography, Journal of Chromatography Library*, vol. 62, Elsevier Science Publishers B.V., Amsterdam.

El Rassi, Z. (ed.) *Electrophoresis*. In this journal, published by Wiley-VCH Verlag GmbH, Weinheim, one can find many special issues related to recent developments in using CE in bioanalysis, CE/MS, microchip separation, or specific application fields.

Foret, F., Krivánková, L., and Boček, P. (eds) (1993) *Capillary Zone Electrophoresis*, VCH, Weinheim.

Guzman, N.A. (ed.) (1993) *Capillary Electrophoresis Technology*, Marcel Dekker, New York.

Guttman, A. (ed.) (2008) *Capillary Gel Electrophoresis and Related Microseparation Techniques*, Elsevier, Amsterdam.

Henry, C.S. (ed.) (2006) *Microchip Capillary Electrophoresis Methods and Protocols*, Humana Press, Totowa.

Khaledi, M.G. (ed.) (1998) *High-Performance Capillary Electrophoresis*, John Wiley & Sons, Inc, New York.

Kuhn, R., and Hofstetter, S. (eds) (1993) *Capillary Electrophoresis: Principles and Practice*, Springer, Heidelberg.

Landers, J.P. (ed.) (1994) *Handbook of Capillary Electrophoresis*, CRC Press, Boca Raton, FL.

Landers, J.P. (ed.) (2008) *Handbook of Capillary and Microchip Electrophoresis and Associated Microtechniques*, CRC Press, Boca Raton, FL.

Li, S.F.Y. (1992) *Capillary Electrophoresis: Principles, Practice and Applications*, Elsevier Science Publishers B.V., Amsterdam.

Lunn, G. (ed.) (2000) *Capillary Electrophoresis Methods for Pharmaceutical Analysis*, John Wiley & Sons, Inc., New York.

Lunte, S.M. and Radzik, D.M. (ed.) (1996) *Pharmaceutical and Biomedical Applications of Capillary Electrophoresis*, Pergamon, Oxford.

Neubert, R. and Ruttinger, H.-H. (eds) (2003) *Affinity Capillary Electrophoresis in Pharmaceutics and Biopharmaceutics*, Marcel Dekker, New York.

Righetti, P.G. (ed.) (1995) *Capillary Electrophoresis in Analytical Biotechnology*, CRC Press, Boca Raton, FL.

Schmitt-Kopplin, Ph. (ed.) (2008) *Capillary Electrophoresis Methods and Protocols*, Springer/Humana Press, Totowa.

Schmitt-Kopplin, Ph. (ed.) (2016) *Capillary Electrophoresis Methods and Protocols*, 2nd edn, Springer, Totowa.

Schmitt-Kopplin, Ph. and Frommberger, M. (2003) Capillary electrophoresis – electrospray ionisation mass spectrometry: 15 years of development and applications. *Electrophoresis*, **24**, 3837–3867.

Strege, M.A. and Lagu, A.L. (eds) (2004) *Capillary Electrophoresis of Proteins and Peptides*, Humana Press, Totowa.

Volpi, N. and Maccari, F. (eds) (2013) *Capillary Electrophoresis of Biomolecules, Methods and Protocols*, Springer.

Wehr, T., Rodriguez-Diaz, R., and Zhu, M. (ed.) (1998) *Capillary Electrophoresis of Proteins*, Marcel Dekker, New York.

Amino Acid Analysis

13

Josef Kellermann

Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany

Many technologies applied in protein chemistry require the knowledge of the exact amount of protein. Amino acid analysis is able to provide much more and, above all, more exact information than colorimetric methods are able to do. Besides the determination of the amount of protein and the relative composition of the amino acids of a peptide or protein, a quantification of free amino acids can also be achieved. The percentage composition of amino acid residues of a protein results in a characteristic profile, sufficient for the identification of a protein in a data base. The concurrent determination of the amino sugars can give initial information about the carbohydrate character of a protein. The amino acid analysis often helps to select the appropriate protease for the specific fragmentation of a protein. Furthermore, the amino acid analysis is utilized for C-terminal sequence analysis.

Protein Determination, Chapter 2

Carbohydrate Analysis, Section 23.3

C-Terminal Sequence Analyis, Section 14.2

Besides its role in protein analytics, amino acid analysis increasingly plays a considerable role in other domains like clinical diagnostic, biomedical research, bioengineering, and foodstuffs industry. For this purpose different technologies have been developed and commercialized. There is still tremendous need to improve these technologies in respect of speed, robustness, reproducibility, and sensitivity. The focus is thereby shifting from the analysis of protein hydrolysates towards the analysis of free amino acids in different biological matrices.

Amino acid analysis was first developed by Stein and Moore in 1948. Their method used 6 M HCl for acid hydrolysis in an oxygen-free sphere at 110 °C for 22 h to release amino acids from proteins. They started to separate the free amino acids using starch columns and detected and quantified the eluted amino acids using a post-column ninhydrin reaction. They soon switched to sulfonated polystyrene beads (Dowex 50). The separation of protein hydrolysates still needed five days, which was half the time necessary when using starch columns as stationary phase. As solvent systems lithium or sodium citrate buffers were used. In principle, this method only differed from the method used today in the amount of analyzed material. In 1958 the separation of an amino acid hydrolysate could already be completed within 24 h. Spackman in the same year published an "instrument to automatically record the color yield of ninhydrin," achieving the quantitative determination of 100 nmol amino acids with an accuracy of 3%. This system turned out to be the classical method for amino acid analysis. For this work Stein and Moore were awarded the Nobel Prize in Chemistry 1972.

Chromatographic Separation Techniques, Chapter 10

Amino acids are small polar molecules that are difficult to separate from each other except by ion exchange chromatography. The separation is carried out with cation-exchange columns that utilize the zwitterionic nature of amino acids properties, whereby, depending on the presence of amino or carboxyl groups, the amino acids can be either negatively or positively charged, depending on the pH of the buffer. Usually amino acids are loaded onto the column in low pH-buffer, which binds the positively charged amino acids to the column resin. Amino acids are then released using increasing ionic strength and pH. Amino acids cannot be easily detected by UV-absorbance or fluorescent and have to be derivatized after elution from the column. During

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

the late-1970s new derivatization techniques allowed for an improvement of the chromatographic as well as detection properties, leading to new systems for amino acid analysis. The introduction of reversed phase HPLC had a crucial impact, decreasing run times to 30 min. Recently, with the development of ultrahigh pressure liquid chromatography (UHPLC) systems, the resolution of the columns has been be increased, resulting in run times for hydrolysates of below 10 min.

Besides optical detection methods coupled to chromatographic systems, mass spectrometry as the detection technique was introduced recently, coupled to separation technologies like chromatography, capillary electrophoresis, or gas chromatography.

13.1 Sample Preparation

The first step during amino acid analysis of peptides or proteins is the cleavage of the peptide bond, which can be achieved by chemical or enzymatically hydrolysis (Figure 13.1).

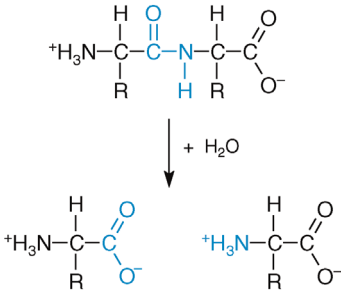


Figure 13.1 Hydrolysis of a peptide bond.

13.1.1 Acidic Hydrolysis

The standard method in protein chemistry, introduced by Moore in 1963, is acidic hydrolysis with constant boiling 6 N HCl (24 h at 110 °C) in an oxygen-free environment. Modifications of this method – such as the use of different acids, elevated temperatures, shortened incubation times, and the addition of diverse scavengers – are necessary to counteract problems, depending on the different hydrolysis behavior of various amino acids. The standard conditions are a compromise between hydrolysis time and the temperature at which some of the amino acids are destroyed partially. This can result in the loss of about 10–40% of serine, threonine, or methionine and 50–100% of cysteine, tryptophan, amino sugars, or phosphorylated amino acids. Asparagine and glutamine are completely desamidated to their acids during hydrolysis. Shortened hydrolysis times indeed improve the yield of the sensitive amino acids but impair the release of amino acids from hydrophobic environment (e.g., Ile-Val or Val-Val). An increase of hydrolysis time to 96 h inverts this ratio. To cope with the hydrolysis behavior of all amino acids, the hydrolysis has to be repeated using several different conditions. The extrapolation of all analyses may lead to a result that is close to the actual amino acid composition of a protein.

With improved sensitivity of the analytical methods impurities in the samples, often introduced by the hydrolysis medium, falsified the results. Therefore, gas phase hydrolysis, where the hydrolysis medium is not in direct contact with the sample, was introduced. Using gas phase hydrolysis and elevated temperatures (e.g., 4 h at 145 °C or even 1.5 h at 165 °C) reduced tremendously hydrolysis time. The combination of different acids is used, such as, for example, propionic acid/HCl (1 : 1) at 160 °C for 15 min or TFA/HCl (1 : 2) at 166 °C for 25 min. Hydrolysis using organic acids like methanesulfonic acid or toluenesulfonic acid results in considerable improvements of tryptophan yields, achieving values of up to 90% without interfering with the yield of the other amino acids.

Especially, tryptophan and methionine are extremely prone to oxidation. Therefore oxygen is removed by alternately evacuating and then filling the hydrolysis vial with inert gas. This is supported by the addition of antioxidants to the hydrolysis medium. As scavengers, phenol (1%), thioglycolic acid (0.1–1%), 2-mercaptoethnaol (0.1%), tryptamino-(3-(2-aminoethyl)-indole), or sodium sulfite are used. For the quantitative determination of cysteine the oxidation of cysteine to cystic acid using performic acid is necessary. Reduction of a protein using thiol followed by alkylation with iodoacetic acid or 4-vinylpyridine results in stable derivatives like *S*-carboxymethyl-L-cysteine or (pyridylethyl)cysteine.

To determine asparagine or glutamine, which are desamidated during hydrolysis, these amino acids have to be rearranged to diaminopropionic acid and butyric acid, respectively, using 1,1-trifluoroacetoxy(iodo)benzene. For the exact determination the values of glutamic acid and aspartic acid with and without pretreatment are subtracted from each other.

The hydrolysis time can be reduced to a few minutes by using gas phase hydrolysis in a microwave oven.

13.1.2 Alkaline Hydrolysis

Alkaline hydrolysis is almost exclusively applied to achieve improved tryptophan yields. 4 M barium, sodium, or lithium hydroxide is used as hydrolysis medium. These strong alkaline media require special reaction vials, as glass would be etched and the released silicate would encourage side reactions. The reaction has to be neutralized following hydrolysis. Barium ions have to be removed by precipitation with carbonate or sulfate, leading to loss of amino acids by adsorption.

13.1.3 Enzymatic Hydrolysis

Enzymatic hydrolysis is also used very rarely. Glutamine and asparagine are not desamidated and can be analyzed directly. This is also a method by which to gently hydrolyze proteins containing sulfated (tyrosine-O-sulfate) and phosphorylated amino acid residues that would be destroyed by acidic hydrolysis. To achieve complete enzymatic hydrolysis several different endo- and exopeptidases with broad specificity have to be used, such as leucine-amino-peptidases, prolidases, subtilisin, papain, and carboxypeptidases. In addition, pronase, a mixture of several unspecific enzymes, is suitable to degrade proteins to single amino acids.

Cleavage of Proteins, Chapter 9

13.2 Free Amino Acids

The analysis of free amino acids is of great importance particularly for physiological samples like plasma or urine as well as in the food industry. Samples often are a complex mixture of nutrients besides proteins, including carbohydrate, fat, vitamins, and minerals and of course free amino acids. Especially high molecular weight substances complicate the analyses as they bind to the stationary phases of columns, reducing the binding capacity of the column and clogging them. Prior analysis, often precipitation, delipidation, extraction, and filtration steps, is necessary.

In contrast to protein hydrolysates with, usually, 18 amino acids (asparagine and glutamine are desamidated) the analysis of physiological samples requires the separation and quantification of up to 50 different components, imposing much higher demands on the analysis system used.

13.3 Liquid Chromatography with Optical Detection Systems

As already mentioned, most analysis systems require a derivatization of the amino acids. This derivatization step can be introduced before (pre-column derivatization) or after (post-column derivatization) the chromatographic separation of the amino acids. The reagents should fulfill the following criteria:

Chromatographic Separation Techniques, Chapter 10

- react with primary and secondary amines;
- result in a quantitative and reproducible reaction;
- each amino acid should form a single stable derivative;
- derivatives should have high UV-absorption or high fluorescent yield;
- reagent and by-products should not be absorbed or disturb chromatography;
- the reaction should take place under mild conditions.

13.3.1 Post-Column Derivatization

During post-column derivatization amino acids were first separated by ion exchange chromatography using a step gradient followed by the addition of the derivatization reagent using a second solvent pump. A reaction coil, with dimensions enabling a correct reaction time, allows the amino acids to react continuously with the reagent. A detector is used to quantify the modified amino acids or corresponding reaction products. Originally, ninhydrin was used for derivatization. To improve sensitivity, fluorescamine or *ortho*-phthaldialdehyde (OPA) can be also used.

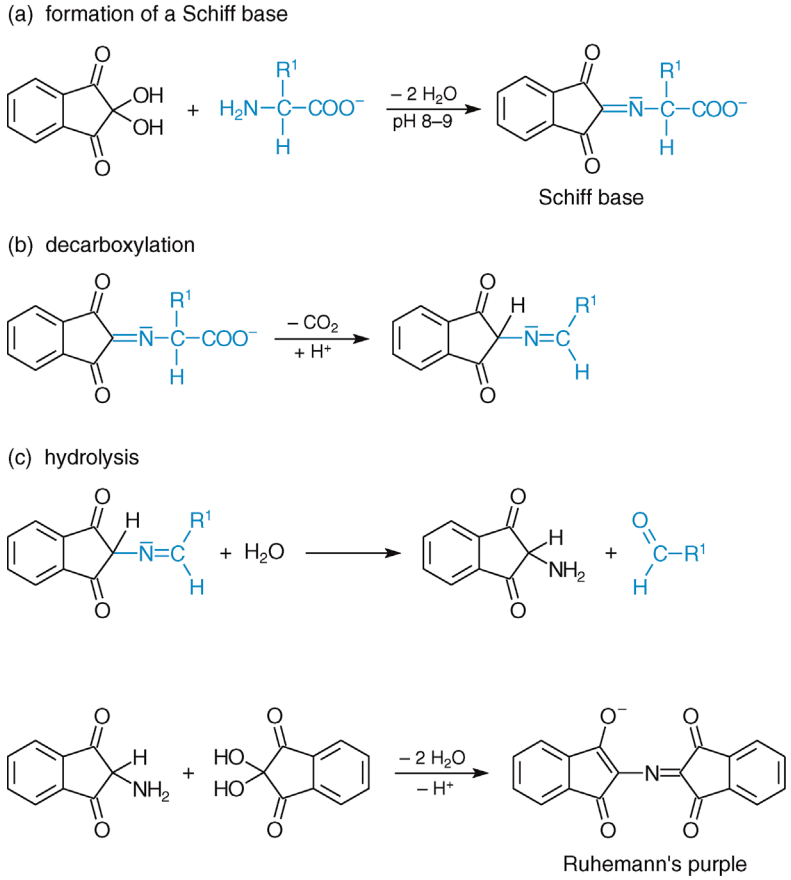


Figure 13.2 Ninhydrin-reaction with primary and secondary amines. The reaction with ninhydrin results in an oxidative decarboxylation of the amino acid; the resulting ammonia and hydrindantin form a purple dye together with another ninhydrin molecule.

Ninhydrin Since the early 1950s, when Stein and Moore developed this method, considerable improvements in speed, sensitivity, and instrumentation have been made, without changing the principle of the method. On passing through the reaction coil, ninhydrin quantitatively reacts with primary and secondary amines (Figure 13.2) at 100–130 °C. Ninhydrin, via the formation of a Schiff base, leads to the decarboxylation of the amino acid. Hydrolysis of the Schiff base of the decarboxylated product affords an aldehyde and a ninhydrin derivative, with the nitrogen of the amine. This ninhydrin derivative, with its middle carbonyl-C-atom, reacts with a second ninhydrin molecule to give a Schiff base, which is deprotonated to a purple dye. The remaining part of the amino acid is not used for detection. The amino acid is identified solely by its retention time during chromatography, while the colored dye is used for quantification. The ring structure of proline and hydroxyproline leads to a different reaction resulting in a yellow dye with an absorption maximum at 440 nm. The reaction does not lead to interfering by-products. The absorption maximum of the reaction product is at 570 nm for primary amines and at 440 nm for secondary amines. The separation usually is performed using a spherical ion exchange resin (10% DVB-crosslinked polystyrene 4 × 150 mm) in citric buffer starting at pH 2. The elution is carried out using a step gradient with increasing ionic strength and increasing pH. (Figure 13.3). The detection limit of this method is around 50 pmol.

Fluorescamine Fluorescamine was used to improve the sensitivity of the method compared to ninhydrin. In alkaline medium fluorescamine reacts with primary amines to give a fluorescent derivative (absorption 475 nm), which can be excited at 390 nm (Figure 13.4). Since fluorescamine is not stable in aqueous solutions and the fluorescent optimum is near pH 9, which is far above the pH for ion exchange chromatography, this method cannot really succeed in amino acid analysis.

***ortho*-Phthaldialdehyde (OPA)** Initially, OPA was used in amino acid analysis almost exclusively as a post-column derivatization reagent. It reacts, similarly to fluorescamine, only with primary amines. In the presence of thiol, fluorescamine reacts with an amino acid to yield 1-alkylthio-2-alkyl-substituted isoindole (Figure 13.5). In contrast to ninhydrin,

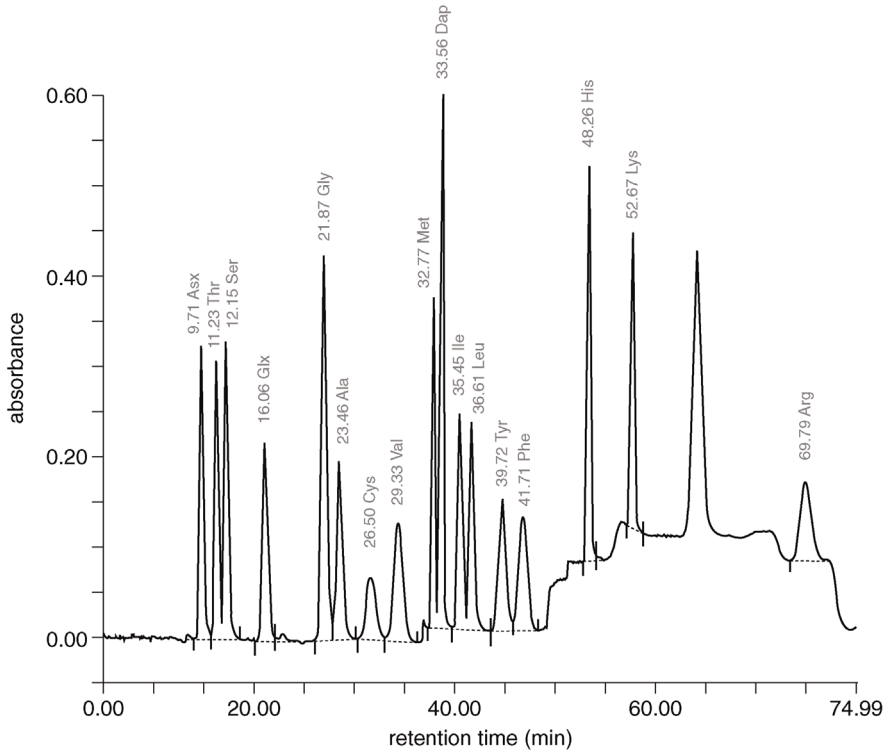


Figure 13.3 Chromatographic separation of a protein hydrolysate standard (1 nmol) by ion exchange chromatography, using a sodium-citrate buffer system. The separation is performed using a three step gradient with increasing salt concentration, increasing pH, and increasing temperature. Detection is carried out at 550 nm.

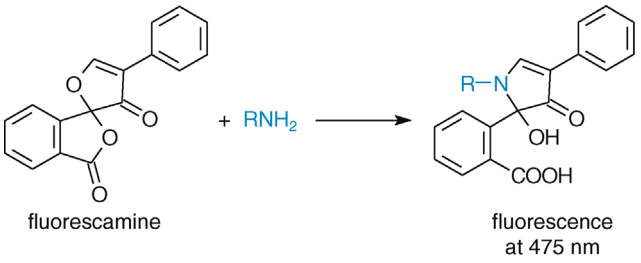


Figure 13.4 Reaction scheme of fluorescamine with primary amines.

fluorescamine as well as OPA form derivatives with amino acids, which allow for an identification of the amino acid residue. Derivatives can be visualized at 230 nm (UV) or at 460 nm (fluorescence, excitation at 330 nm). Secondary amines do not react with OPA. They have to be oxidized using NaOCl or chloramine T prior to derivatization. These reagents can be continuously added to the buffer during chromatography. The detection limit is about 10 pmol (fluorescence). As OPA is not as effective as ninhydrin it is nowadays mainly used for pre-column derivatization.

13.3.2 Pre-column Derivatization

The development of high-performance liquid chromatography (HPLC) – especially reversed phase chromatography – allowed for the development of new derivatization reagents, which changed the chromatographic behavior of amino acids significantly. Polar amino acids coupled to aromatic reagents become much more hydrophobic compounds, which can be easily separated by reversed phase chromatography. Modern HPLC systems and column material nowadays allow for the separation of all amino acid residues in less than 15 min. The introduction of a chromophore or

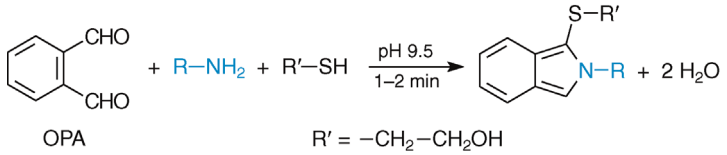


Figure 13.5 Reaction scheme of *O*-phthaldialdehyde with primary amines. Formation of an isoindole-derivative during the reaction of OPA with primary amines in the presence of a reducing agent (β-mercaptoethanol).

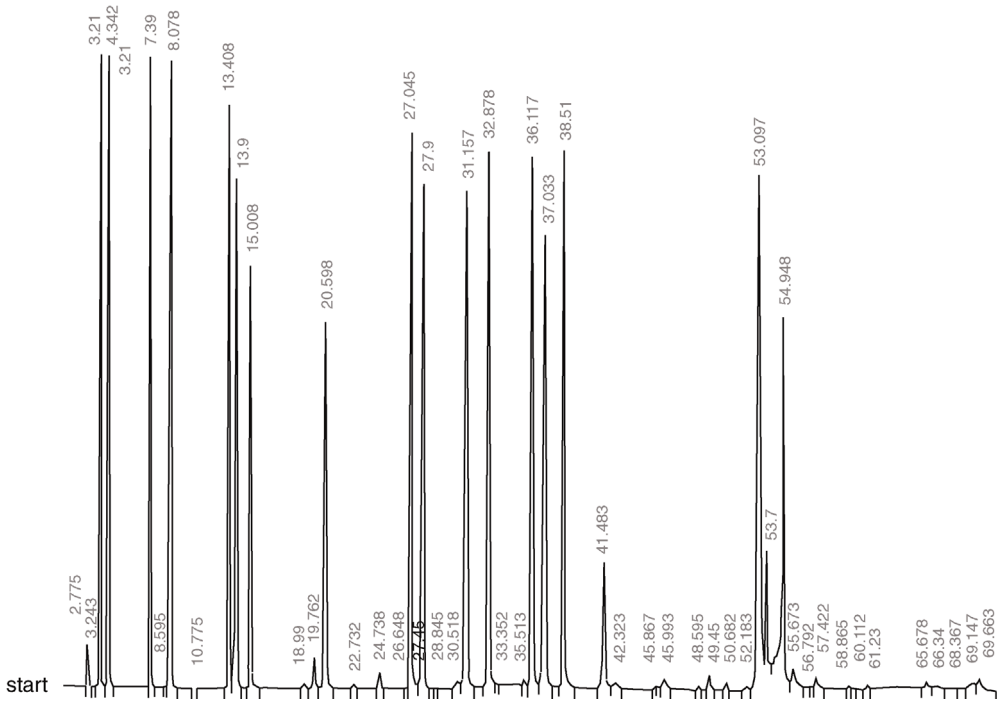


Figure 13.6 Chromatographic separation of a protein hydrolysate, reacted with OPA, by reversed phase chromatography using a C_{18}-column (Shandon 250 mm × 4 mm); buffer A: 10 mM sodium-phosphate, pH 7.2; buffer B: acetonitrile. Flow rate: 1 ml min^{-1}.

fluorophore additionally improves the sensitivity of the method, reaching detection limits down to 50 fmol for some reagents. In practice, this sensitivity range is not realized, since minimal amounts of impurities distort the results, making them useless. Therefore, the sample preparation, even more so than the sensitivity, is the limiting factor of the analysis.

The different stability of the various amino acid derivatives makes short and exactly defined conditions necessary, from derivatization steps to detection, to quantitatively identify even labile amino acid derivatives. This can only be achieved using a completely automated method, including the derivatization step.

***ortho*-Phthaldialdehyde (OPA)** OPA is used in pre- as well as in post-column derivatization (Figure 13.5). Several different thiols are used during derivatization (2-mercaptoethanol, ethanethiol, 3-mercaptopropionic acid); they are responsible for the hydrophobic character and stability of the single derivatives. According to the reagent used different chromatographic parameters have to be chosen, for example, stationary phase and elution buffer. Usually, sodium phosphate buffers, pH 7.2 (Figure 13.6) and an acetonitrile gradient is used. The detection limit is about 10 pmol (UV) or 200 fmol (fluorescence).

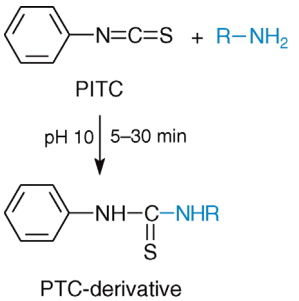


Figure 13.7 Reaction of phenyl isothiocyanate with primary and secondary amines (PITC) to form a phenyl-thiocarbamoyl derivative (PTC).

Phenyl Isothiocyanate (PITC) PITC (phenyl isothiocyanate) is a well-established reagent in protein chemistry since its introduction in Edman degradation. PITC reacts with primary and secondary amines under alkaline conditions within 20 min. The resulting phenylthiocarbamyl (PTC) derivatives of the amino acids (Figure 13.7) are stable and do not interfere with reaction by-products during chromatography. The absorption maximum is around 245 nm with a detection limit of 1 pmol.

9-Fluorenylmethoxycarbonyl (FMOC) Chloride FMOC (9-fluorenylmethoxycarbonyl) chloride is well known based on its role as a protecting group during peptide synthesis. In 1983, FMOC was first described as a derivatization reagent in amino acid analysis (Figure 13.8). The fast reaction with primary as well as secondary amines at pH 4.2 leads to stable derivatives. FMOC chloride hydrolyses under the applied conditions and has to be removed from the reaction mixture as the reaction products interfere with co-eluting amino acid residues. This additional extraction step always leads to unspecific loss of amino acids. The absorbing wavelengths of the

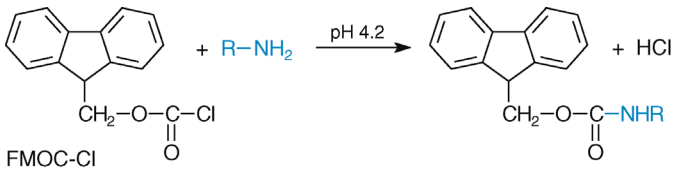


Figure 13.8 Reaction of 9-fluorenylmethoxycarbonyl (FMOC) chloride with primary and secondary amines.

amino acids are at 260 nm (UV) and 305 nm (fluorescence, excitation at 266 nm). The detection limit is around 50 fmol.

Dabsyl Chloride (DABS-Cl) 4-Dimethylaminoazobenzene-4-sulfonyl chloride (DABS-Cl) was first described for amino acid analysis in 1975. Derivatization (Figure 13.9) of primary as well as of secondary amines is performed at 70 °C and pH 9.0 within 15 min. The derivatives are stable for weeks and absorb at 436 nm. This absorption in the visible range leads to a very stable baseline during chromatography, independent of the chromatographic system used. The detection limit is 1 pmol. However, one disadvantage is that it is necessary to know roughly the amount of protein used, since it is important to employ a fourfold excess of the reagent.

Dansyl Chloride 1-Dimethylaminonaphthalene-5-sulfonyl chloride (dansyl chloride) was initially used in amino-terminal sequence determination of proteins and peptides and was first described in amino acid analysis in 1981. It reacts with primary and secondary amines (Figure 13.10), resulting in strongly fluorescent derivatives. Since the reaction is incomplete and results in various by-products, this method has scarcely been of any importance in amino acid analysis.

6-Aminoquinoyl-*N*-hydroxysuccinimidyl Carbamate (ACQ) A commercially available analysis kit is based on the derivatization of primary and secondary amines with 6-aminoquinoyl-*N*-hydroxysuccinimidyl carbamate (ACQ, Figure 13.11). The resulting derivatives are stable for several days at room temperature. The derivatization step, which often is critical during amino acid analysis and can only be achieved reproducibly by automation, can be performed without problems manually using the ACQ method. Excess reagent hydrolyzes to aminoquinoline (AMQ), *N*-hydroxysuccinimide (NHS), and carbon dioxide. The main by-product AMQ is slightly fluorescent at 395 nm but does not interfere with amino acids during reversed phase chromatography. The detection is carried out at 395 nm (excitation 250 nm) and can also be done at 248 nm (UV) with lower sensitivity. The detection limit is about 100 fmol.

Chiral Reagents to Identify Enantiomer Amino Acids The analysis of enantiomeric amino acids is particularly important during quality control of amino acids for peptide synthesis, control

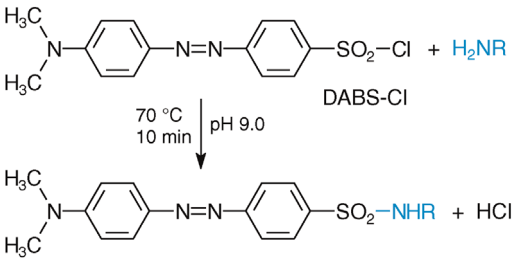


Figure 13.9 Reaction of dabsyl chloride (DABS) with primary and secondary amines.

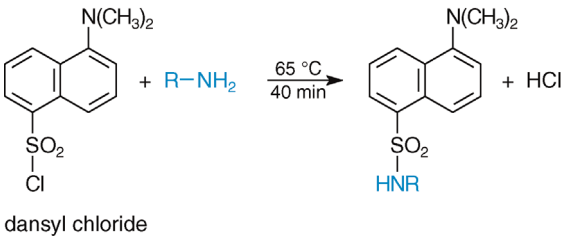


Figure 13.10 Reaction of dansyl chloride with primary and secondary amines.

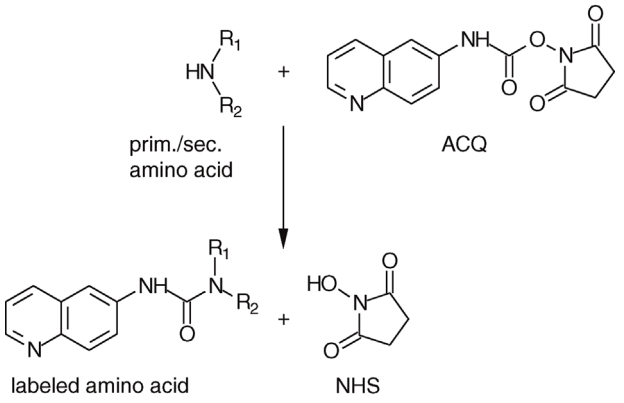


Figure 13.11 Reaction of 6-aminoquinoyl-*N*-hydroxysuccimidyl carbamate (ACQ) with primary and secondary amines.

of peptide pharmaceuticals, and in the food industry. Enantiomers are analyzed by the formation of diastereomeric complexes. For the formation of diastereomers with D- as well as L-amino acids it is necessary to use optically clean reagents like (+)-1-(9-fluorenyl)ethyl chloroformate (FLEC) (Figure 13.12). The reaction is similar to the reaction of FMOC and can be performed using borate buffer, pH 6.8 at room temperature. Chromatographic separation of the derivatives is achieved using C_8 or C_{18} reversed phase columns. The derivatives are detected using a UV-detector at 254 nm or a florescence detector at 315 nm (excitation at 260 nm). Alternatively, OPA can be used together with thiols to form isoindolyl derivatives for the analysis of enantiomers.

Another method employed to analyze enantiomers is their HPLC separation using chiral stationary phases. Reversible diastereomeric complexes are formed between the chiral stationary phase and the adsorbed derivative (Figure 13.13).

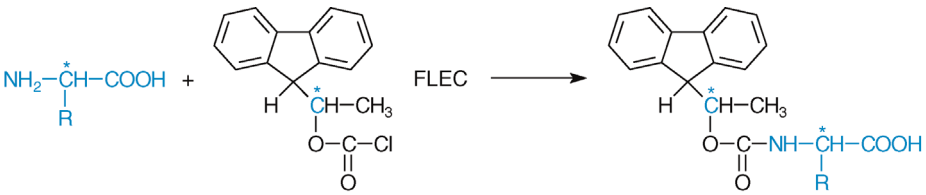


Figure 13.12 Reaction of chiral (+)-1-(9-fluorenylethyl)-chloroformate with D- and L-amino acids leads to diastereomeric complexes.

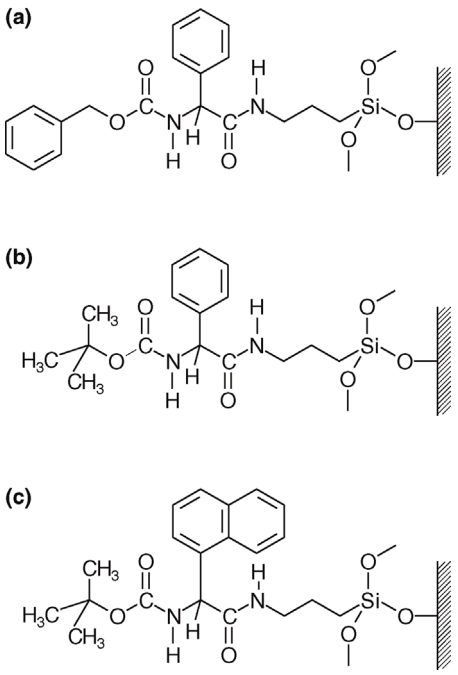


Figure 13.13 Structures of chiral stationary phases coupled via silanol: (a) (D-phenylglycine)$_2$-NH(CH$_2$)$_3$-Si(OC$_2$H$_5$)$_2$; (b) Boc-D-phenylglycine-NH(CH$_2$)$_3$-Si(OC$_2$H$_5$)$_2$; (c) Boc-L-(1-naphthylglycine-NH(CH$_2$)$_3$-Si(OC$_2$H$_5$)$_2$.

13.4 Amino Acid Analysis using Mass Spectrometry

Unlike optical detection using photometry, another separation dimension is introduced when mass spectrometry is employed as detecting technology, since the analytes additionally are separated according to their mass. Therefore, high resolution pre-separation using chromatography, electrophoresis, or gas chromatography is not necessary. The time needed for analysis of a single sample can be reduced significantly, thereby increasing sample throughput. But, the technical effort is dramatically higher when using mass spectrometry, which can only be justified by high sample throughput. Therefore, only the few methods described here are applied routinely.

Mass Spectrometry, Chapter 15

Ion Pair-LC-MS-MS Amino acids can be separated without derivatization using ion pair chromatography at C_{18}-reversed phase columns. Using volatile buffers, an electrospray mass spectrometer directly can be coupled to these chromatographic systems. The amino acids are quantified in MRM mode and the addition of stable isotopes to the sample. Since the MRM mode introduces a third separation dimension, the analysis time can be reduced to 20 min for more than 70 amino acids from physiological samples.

SRM and MRM Analysis, Section 15.2.6, 39.5.7

HILIC-MS Hydrophilic interaction uses polar stationary phases like silica gel-, amide-, hydroxyl-, cyano-, or amino-column material. Polar analytes are bound on running the column with organic solvents like acetonitrile. The amino acids are then eluted with increasing aqueous buffers. The detection is carried out as described above using a MS or MS-MS as detection system.

HILIC, Section 10.4.4.

CE-MS Since amino acids are charged molecules they can be separated without derivatization using capillary electrophoresis. Detection is performed by MS methods as described above. The disadvantage of this method is the small sample volume that can be used in capillary electrophoresis, resulting in lower sensitivity.

Capillary Electrophoresis, Chapter 12

GC-MS Gas chromatography is an ideal separation technology when used together with mass spectrometry. However, amino acids have to be modified prior to separation. Silanization is often used, replacing active hydrogen with alkylsilyl molecules. Often trimethylsilyl-derivatives are prepared by reaction with *N,O*-bis-(trimethylsilyl)trifluoroacetamide (BSTFA) or *N*-methyl (trimethylsilyl)trifluoroacetamide (MSTFA), but unfortunately these derivatives are not stable. Alternatively, amino acids can be acylated or esterified, using anhydride/alcohol combinations like pentafluoropropyl anhydride and isopropanol.

aTRAQ-LC-MS-MS Based on iTRAQ chemistry, used for relative quantification of peptides during proteomics experiments, a method was introduced to react the amino group with NHS-esters coupled to a linker molecule. The tags also contain two different cleavable reporter ions, one for the sample and one for the standard, which are detected by fragmentation in a tandem mass spectrometer. Reversed-phase liquid chromatography of the labeled amino acids is performed prior to mass spectrometric analysis.

Stable Isotopic Labeling (i-TRAQ), Section 39.6.2

Direct Infusion MS-MS Direct infusion MS-MS routinely is used in clinical diagnostics. Screening of blood and urine from newborn children to analyze metabolic disorders is performed in a high throughput manner. Physiological samples are collected on filter sheets that already contain isotopic labeled standard amino acids for absolute quantification of the sample. An area of defined size is cut and the amino acids are extracted using methanol. The addition of hydrochloric acid in *n*-butanol converts the amino acids into their corresponding butyl esters, which can be analyzed by MS-MS without further separation. The missing separation step allows for extremely high sample throughput with the restriction that isobaric amino acids like isoleucine and leucine or alanine and sarcosine cannot be distinguished. Using a high resolution mass spectrometer like FT-ICR or Orbitrap, amino acids with almost identical nominal mass can be distinguished.

Table 13.1 Reagents used for amino acid analysis.

	Detection method	Limit of detection	Analysis time (min)	Use
Ninhydrin	UV 570 and 440 nm	50 pmol	80	Primary/secondary amino acids.
PITC	UV 245 nm	10 pmol	30	Primary/secondary amino acids
Fluorescamine	Fluorescence (390/475 nm)		90	Primary amino acids
OPA	UV 230 nm; fluorescence (330/460nm)	10 pmol/200 fmol	30	Primary amino acids
FMOC	Fluorescence (266/305 nm)	50 fmol	30	Primary/secondary amino acids
Dabsyl-Cl	UV 436 nm	1 pmol	30	Primary/secondary amino acids
ACQ	UV 248 nm; fluorescence (250/395 nm)	100 fmol	35	Primary/secondary amino acids

13.5 Summary

Despite exact calibration and the introduction of internal standards it is necessary to evaluate the results of an amino acid analysis thoroughly. There are only a few amino acids like alanine, phenylalanine, or leucine, where low amounts can be analyzed reliably. The values of serine and threonine are always to low, since the amino acids are partially destroyed during hydrolysis. Cysteine and tryptophan are destroyed nearly completely and the values for valine and isoleucine are often too low since the hydrolysis of peptide bonds involving these two amino acids is incomplete. Methionine, as well as tyrosine, is extremely sensitive towards oxidation, often resulting also in low values (Section 13.1.1). To overcome these problems and to achieve

Table 13.2 Overview of different technologies for amino acid analysis.

Method	Advantages	Disadvantages
LC-Methods with optical detection	Methods well established for a long time, highly reproducible, inexpensive equipment, good linearity	Protein removal necessary for free aa, derivatization necessary, no specificity for analytes, co-eluting analytes cannot be distinguished
LC-MS	Fast separation, high resolution	Protein removal necessary for free aa
IP-LC-MS-MS	No derivatization step, high number of analytes, high resolution of polar analytes	Protein removal necessary for free aa, ion suppression
HILIC-MS	No derivatization step, well suited for polar analytes	Protein removal necessary for free aa, poor reproducibility, ion suppression
CE-MS	No derivatization step, small amount of sample	Protein removal necessary for free aa, small injection volume
GC-MS	Robust, highly reproducible, high resolution, fast separation,	Derivatization necessary, derivatives not table
iTRAQ-MS	Fast separation, internal standard for all analytes	Protein removal necessary for free aa, poor recovery for sulfur-containing aa, poor reproducibility
Direct infusion-MS-MS	No separation necessary, high throughput	extraction and derivatization necessary, isobaric aa cannot be distinguished

reliable results and exact quantification despite the different hydrolysis behavior it is necessary to hydrolyze samples at different temperatures and to extrapolate the achieved values for each amino acid. Glycine values often are too high, which is caused by contamination. Some reagents (OPA or fluorescamine) do not react with secondary amines at all. However, nowadays, amino acid analysis is routinely reliable quantitatively down to about 10 pmol amino acid with an error of about 10%. The limit of detection even is in the low fmol range (Table 13.1). The major part of the error arises from hydrolysis of the protein. The analytical part can be performed with errors of less than 2%. All pre- as well as post-column derivatization methods suffer from their long analysis time. These can be dramatically shortened when mass spectrometry is used as detection technology (Table 13.2) since with its high selectivity and specificity another separation dimension is introduced. With the addition of isotopic labeled amino acids as internal standard and quantification using the MRM mode, mass spectrometry based amino acid analysis becomes a fast, sensitive, and reliable method, without a derivatization step, that might replace standard technologies when high throughput justifies more expensive equipment.

Further Reading

Alterman, M.A. and Hunziker, P. (eds) (2009) *Amino Acid Analysis. Methods and Protocols,* Methods in Molecular Biology, vol. 828 Humana Press. Totawa, NJ.

Barrett, G.C. (1985) *Chemistry and Biochemistry of the Amino Acids*, Chapman & Hall, London.

Blackburn, S. (1978) *Amino Acid Determination, Methods and Techniques*, Marcel Dekker, New York.

Cooper, C., Packer, N., and Williams K. (eds) (2001) *Amino Acid Analysis Protocols*, Humana Press, Totowa, NJ, 265 pp

Kaspar, H., Dettmer, K., Gronwald, W., and Oefner, P.J. (2009) Advances in amino acid analysis. *Anal. Bioanal. Chem.*, **393**, 445–452.

Rutherford, S.H. and Gilani, G.S. (2009) Amino acid analysis, in *Current Protocols in Protein Science*, John Wiley & Sons, Inc., Hoboken, 11.9.1–11.9.37.

Protein Sequence Analysis

14

Friedrich Lottspeich

Peter-Dörfler-Straße 4a, 82131 Stockdorf, Germany

By 1940 it was already agreed that proteins consist of amino acids and that the amino acids are linked via the so-called peptide bond (Figure 14.1). It was known that the resulting chain-like molecules bear at one end, referred to as the N-terminal end, a free amino group and at the other end, the C-terminal end, a free carboxyl group. It was not agreed, however, at that time whether a particular protein is composed of a mixture of different polymers – each consisting of a defined number and type of amino acids but with different sequences – or of a single species of molecules having a unique defined amino acid sequence. This question was answered only in 1953, at least for small proteins, when Sanger and coworkers elucidated the complete amino acid sequence of the peptide hormone insulin. Sanger used reagents that mark specifically the terminal amino acid residues. For example, 1-fluoro-2,4-dinitrobenzene, the so-called Sanger's reagent, specifically reacts with the free amino group of the N-terminal end of a peptide chain. After complete hydrolytic cleavage of the peptide into amino acids, the marked (terminal) amino acid was isolated and identified by chromatographic techniques. Unfortunately, in this type of analysis one could only find information on the N-terminal end of the peptide chain, since the peptide chain had to be destroyed in order to isolate and identify the labeled amino acid. To obtain sequence information from larger peptides or proteins it was necessary to cleave them via partial hydrolysis or enzymatic digestion into small fragments and then determine the N-terminal and amino acid composition of each fragment with the methods proposed by Sanger. In this way, by analyzing many small fragments the entire sequence of 51 amino acid residues was determined and thus it was shown for the first time that insulin, a small protein, has only one defined amino acid sequence. This extremely tedious work stretched over ten years and consumed about a hundred grams of insulin. Sanger was awarded the Nobel Prize in Chemistry 1958.

Cleavage of Proteins, Chapter 9

Amino Acid Analysis, Chapter 13

A more efficient method for the determination of peptide sequences was published in 1950 by the Swedish scientist Pehr Edman and by in the mid-1950s it had displaced the Sanger method completely. In his work on the sequential degradation of proteins and peptides Edman describes a reaction cascade that has become known as Edman degradation. Edman developed not only a

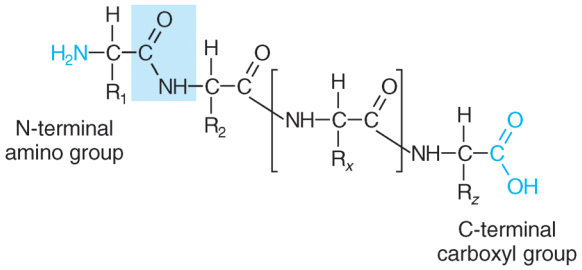


Figure 14.1 General structure of a peptide.

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

new reagent and the chemical mechanism of a cyclic reaction that can cleave an amino acid at a time from the N-terminal end of the peptide chain, but also gave detailed experimental guidance for the identification and quantification of the reaction products. By this he introduced a completely elaborated system for N-terminal amino acid sequence analysis. This has contributed significantly to the rapid acceptance and success of the method. Using Edman degradation G. Braunitzer could determine in 1961 the amino acid sequence of the first larger protein, of human hemoglobin. This made it clear that not only peptide hormones but also proteins have a characteristic and defined sequence. The elucidation of the amino acid sequence of novel proteins and peptides is still performed by Edman degradation. Since the introduction of the method the sensitivity has been increased by a factor of 10^3. Today amino acid sequences may be obtained from a few picomoles of protein. However, the Edman degradation also has inherent limitations (see below), which restrict the information in a single N-terminal sequence analysis to about 30–60 amino acid residues. Therefore, with larger peptides or proteins it is necessary to produce fragments, and to separate these chromatographically or electrophoretically and then subject the individual fragments to sequence analysis again.

The results of Sanger and Edman had an enormous impact on biochemistry as a whole, as it was so clearly shown that proteins have defined amino acid sequences. Since the amino acid sequence (primary structure) in principle provides the basis for the folding and thus of the spatial structure (tertiary structure) of the protein, it is also ultimately responsible for the function of the protein. Knowledge of the amino acid sequence, therefore, is extremely important in understanding the function at the molecular level. This applies even more so to other methods of protein structure determination, such as X-ray crystallography or NMR spectroscopy, which need a known amino acid sequence on which to base the interpretation of their data.

Another example of the importance of sequence analysis in understanding structure–function relationships is the homology comparisons of isoenzymes or functionally equivalent proteins from different species. Amino acid residues that are important for the function of the protein and therefore are conserved in evolution can be detected. Only 20 years ago, sequence analysis was virtually the only way to complete the primary structure determination of proteins.

The situation has changed dramatically due to the enormous development of molecular biological techniques and mass spectrometry. The ease and speed with which the primary structural information of proteins can be obtained via molecular biological techniques today has changed the strategy used to elucidate protein primary structures. The nucleotide sequence information of entire genomes is translated to protein sequences and is stored in protein databases. Today, almost always, a protein will be identified and characterized by sequences of short fragments (i.e., peptides) obtained after enzymatic cleavage, mainly by mass spectrometry techniques but sometimes still by Edman sequencing. Using these peptide sequences databases are searched. This approach is called “bottom up” protein analysis. However, it has some fundamental limitations since it assumes that a peptide is a surrogate for a protein, which in many cases is certainly not the case. A certain peptide may be derived from different proteins originating from completely different DNA sequences or from different splicing products. Database analysis indicates that this is not a rare event. Furthermore, a peptide may stem simply from a common sequence in isoforms or processed (e.g., truncated) forms of a protein or protein family. Furthermore, the numerous post-translational processing and modification events lead to a plethora of distinguishable protein molecules, which have a great part of the sequence in common but in fact are different on the molecular level and perhaps also on the functional level. Especially in quantitative analyses this may cause severe problems and the results of all kinds of bottom up approaches are to be queried and extensively validated by other techniques (e.g., total protein mass determination, total protein sequence analysis, purity control, analysis of heterogeneity, etc.).

Proteome Analysis, Chapter 39

Nevertheless, currently the vast bulk of amino acid sequence analysis is the elucidation of partial sequences of unknown proteins, which then becomes the design of oligonucleotide probes to isolate the cDNA. However, notably, a great risk arises from the current practice of almost exclusive use of molecular biological techniques to create complete protein sequences: Many post-translational modifications that will largely determine the properties and functions of proteins are, significantly, not encoded by the amino acid sequence and therefore cannot be

detected, or even suspected at the DNA level. Therefore, for complete characterization and structure determination of a protein more complementary analyses for the detection of post-translational modifications are essential. Here, all the elaborated protein chemistry techniques of protein purification, cleavages fractionation, and the most efficient analytical procedures are required. In the analysis of post-translational modifications additional to amino acid sequence analysis, mass spectrometry has established itself as an extremely efficient tool. By accurate mass determination together with a known DNA sequence post-translational modification can be suspected or excluded in many cases.

Mass spectrometry, which can provide amino acid sequence information, is now a complementary and fast alternative to the classic sequence analysis. It is used mainly for small peptides of 15–20 amino acid residues. There are limitations with larger proteins and sequences from organisms with unknown genome.

Mass Spectrometry, Chapter 15

The progress made in recent years in data processing programs allows, relatively easily, automatic analysis of mass spectrometric data and – at least with known amino acid sequences – fast and secure localization of post-translational modifications. Especially for sequencing of proteins from organisms with an unknown genome and for the determination of the N-terminus of a protein or of protein fragments the classical amino acid sequence analysis based on the Edman degradation is still the method of choice.

14.1 N-Terminal Sequence Analysis: The Edman Degradation

14.1.1 Reactions of the Edman Degradation

The Edman degradation is a cyclic process, where in each reaction cycle the terminal (N-terminal) amino acid residue of the peptide chain is cleaved off and is identified. The reaction consists of three mutually well-defined steps: coupling, cleavage, and conversion. In the first step, the coupling, the Edman reagent phenyl isothiocyanate (PITC) (Figure 14.2) is bound to the free N-terminal amino group of the peptide chain. This reaction proceeds almost completely at temperatures of 40–55 °C and reaction times of 15–30 min. Thus, a disubstituted thiourea, the phenylthiocarbamoyl (PTC)-peptide, is formed. The addition of PITC is only possible to unprotonated amino groups, therefore, the pH must be maintained in this reaction by an alkaline buffer at about 9. An even higher pH would further improve the reaction rate, but also accelerates an important side reaction: the alkali-catalyzed hydrolysis of PITC to aniline. The resulting aniline reacts with its free amino group with PITC to give diphenylthiourea (DPTU), the only significant by-product of the Edman degradation (Figure 14.3). With a nonpolar solvent (e.g., ethyl acetate), in which the protein as a hydrophilic molecule is not soluble, the excess of reagent and a large part of the DPTU, both relatively hydrophobic components and readily soluble in ethyl acetate, are separated from the PTC-peptide.

Chemical Modification of Proteins, Chapter 6

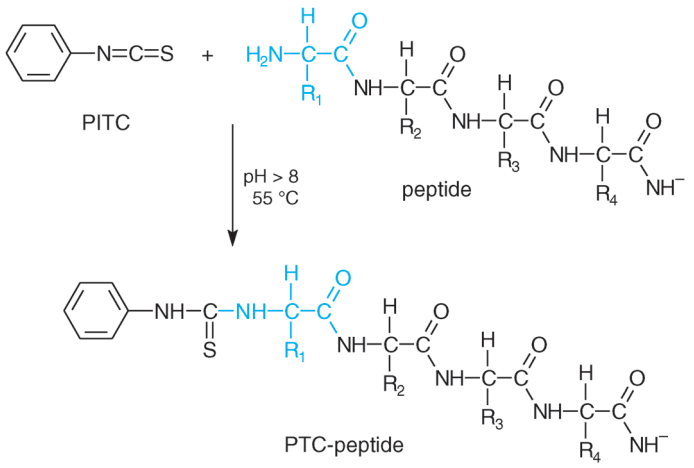


Figure 14.2 Coupling reaction of Edman degradation. Phenyl isothiocyanate (PITC) couples to the free amino group of a peptide to form a phenylthiocarbamyl peptide (PTC-peptide).

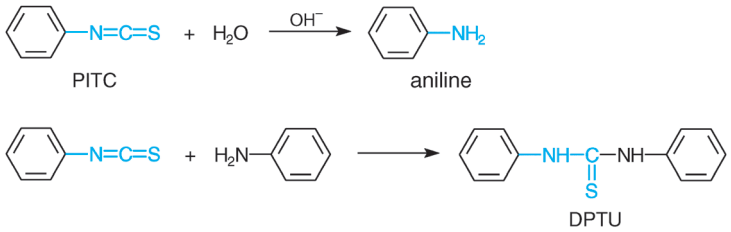


Figure 14.3 Generation of diphenylthiourea (DPTU) during Edman degradation. Phenyl isothiocyanate (PITC) is hydrolyzed to aniline, which reacts with another molecule of PITC to give DPTU.

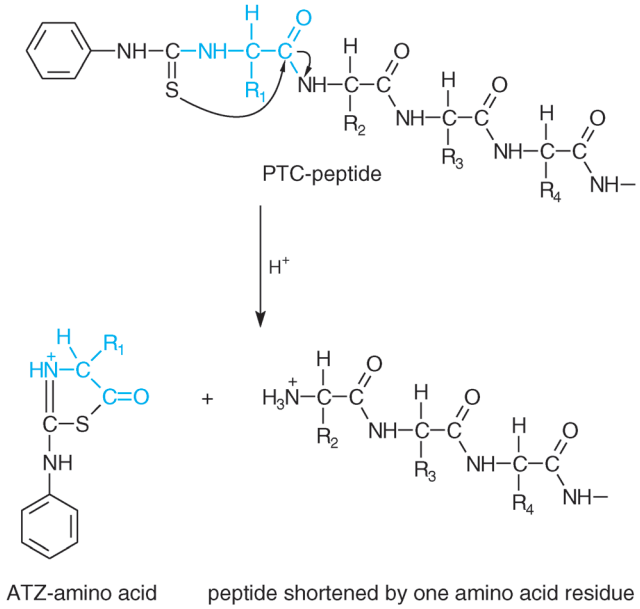


Figure 14.4 Cleavage reaction of Edman degradation. Under acidic, water-free conditions the sulfur performs a nucleophilic attack on the carbonyl group of the first peptide bond. A relatively unstable anilinothiazolinone (ATZ) amino acid and the peptide shortened by one amino acid residue are formed. This shortened peptide exhibits, like the starting peptide, a free amino group.

In the cleavage reaction of the Edman degradation, the dried PTC-peptide is treated with an anhydrous acid (e.g., trifluoroacetic acid). Here, by a nucleophilic attack of the sulfur at the carbonyl group of the first peptide bond, the first amino is cleaved off as a heterocyclic derivative, an anilinothiazolinone (ATZ) amino acid (Figure 14.4). Here the importance of PITC used for the coupling becomes clear, since only the sulfur is nucleophilic enough to lead to ring formation. If there is an oxygen in place of sulfur (i.e., an isocyanate is used as coupling reagent), the reagent can be coupled to the amino group of the peptide as well, but is not able to form a ring and therefore also not able to cleave off the amino acid. Therefore, any sulfur–oxygen exchange in the PITC and the PTC-peptide must be prevented. This is done by using an inert gas atmosphere during the entire Edman degradation.

After evaporation of most of the volatile acid, the small and relatively hydrophobic ATZ-amino acid, which differs significantly in solubility behavior from the hydrophilic peptide, is extracted with a hydrophobic solvent (chlorobutane or ethyl acetate). The residual peptide, now shortened by one amino acid, is dried and can be subjected to further reaction cycles, where again each terminal amino acid is cleaved.

The ATZ amino acid is chemically unstable and has to be changed in a separate step, the conversion, into a more stable derivative, the phenylthiohydantoin (PTH) amino acid (Figure 14.5). During conversion the unstable ring structure of the ATZ-amino acid is opened with aqueous acid and rearranges under elevated temperature to the thermodynamically more stable PTH-amino acid. The PTH-amino acids are usually identified and quantified by chromatography in comparison to a reference sample that contains the PTH derivatives of all known amino acids (Figure 14.6).

14.1.2 Identification of the Amino Acids

The PTH-amino acids show characteristic UV spectra with an absorption maximum at 269 nm and a specific molar absorption coefficient at 269 nm of $\varepsilon = 33\,000\ \text{mol}^{-1}$. They can be

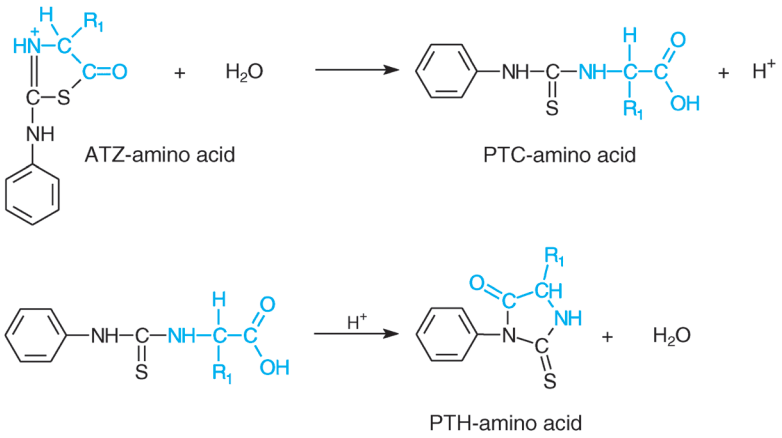


Figure 14.5 Conversion reaction of the Edman degradation. The ATZ-amino acid is hydrolyzed to the phenylthiocarbamoyl (PTC) amino acid. The ATZ amino acid is converted by an acid catalyzed rearrangement into the stable phenylthiohydantoin (PTH) amino acid.

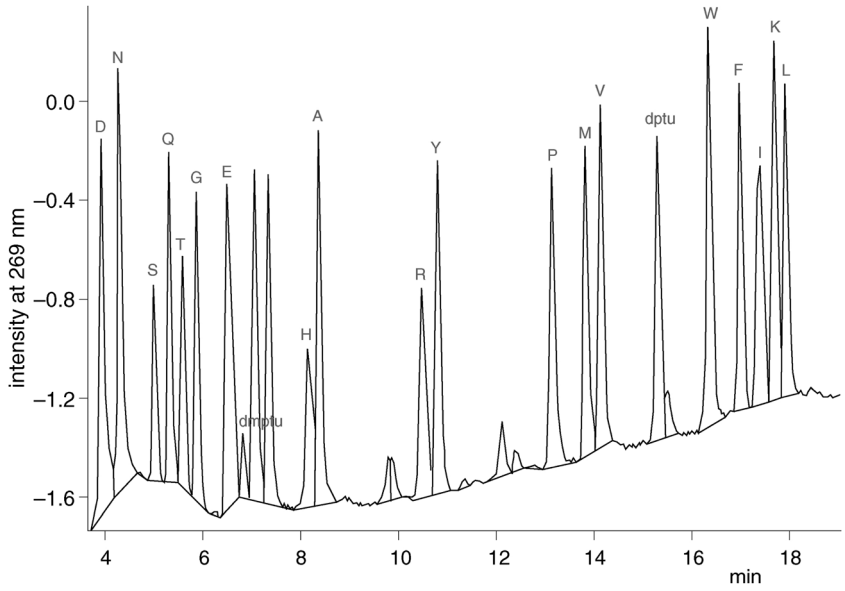


Figure 14.6 PTH-amino acid analysis. The PTH-derivatives of the naturally existing amino acids can be separated by a RP-HPLC system. The PTH-amino acid is identified by comparison of the retention time with the retention time of a standard amino acid mixture.

identified with a detection limit in the femtomole range using microbore reversed phase-HPLC systems. The limit of quantification for most PTH-amino acids using chromatographic methods is, today, below 1 pmol. Since the initial publication in 1950 there has been a constant effort to change the chemical reactions described by Edman in order to improve the detection sensitivity for the eliminated amino acid derivatives. Above all, various fluorescent isothiocyanates, with detection limits in the lower femtomole range, have been proposed as coupling reagents. But none of these reagents has prevailed since they lack quantitative reaction yields either in the coupling or in the cleavage step. In addition, separation of the fluorescent amino acid derivatives, which due to the bulky fluorophore behave chromatographically in similar fashion, poses major problems.

14.1.3 Quality of Edman Degradation: the Repetitive Yield

The quality of the sequence degradation is expressed objectively by the repetitive yield. It describes the overall yield of a degradation step in the Edman degradation. As can be seen from Figure 14.7 the portion of completely degraded peptide chains is reduced in each degradation step. After several degradation steps, especially with low repetitive yields, the number of longer molecules (caused by incomplete reactions) accumulates and may even exceed the quantity of

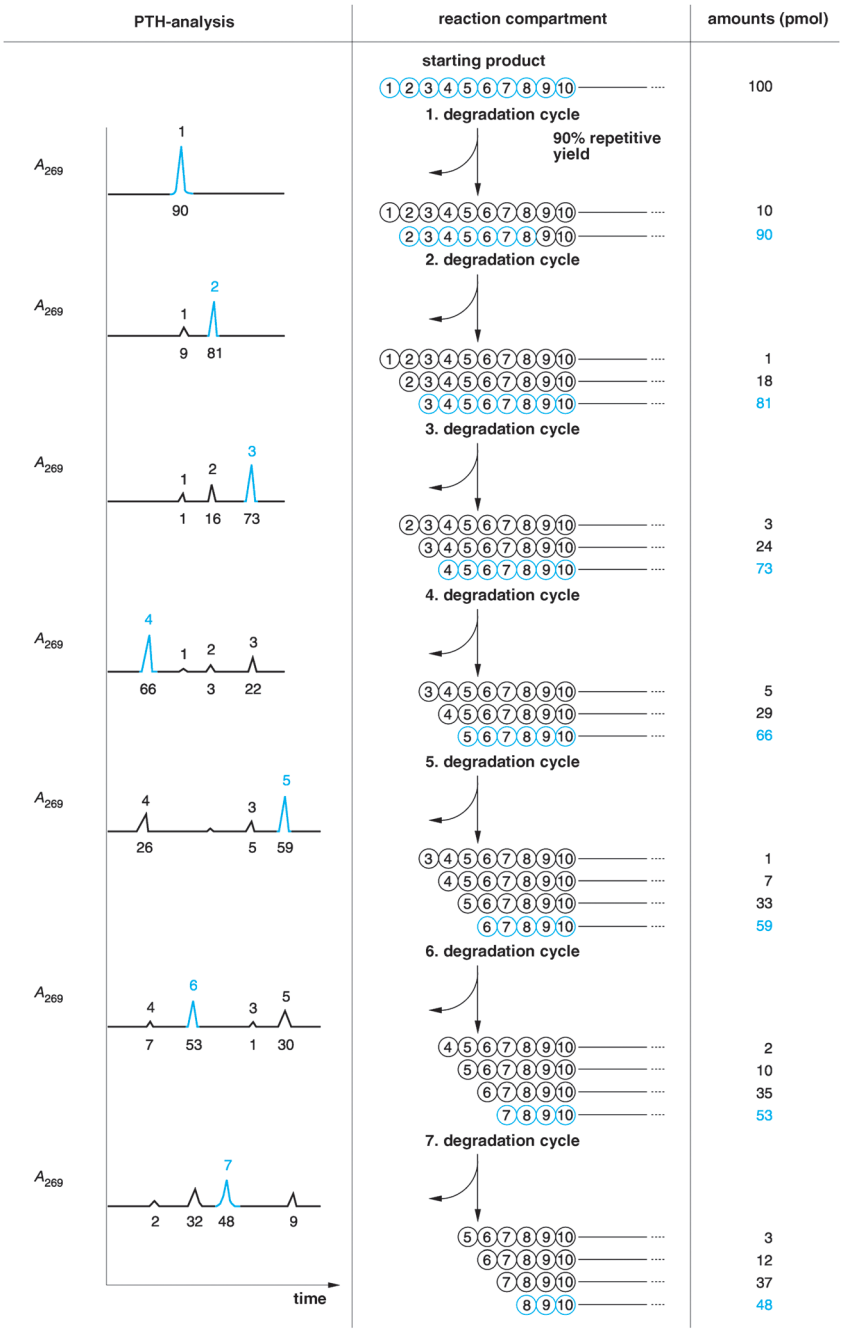


Figure 14.7 Repetitive yield in Edman degradation. A peptide (100 pmol) is sequenced with a repetitive yield of 90%. Obviously, the amount of the expected and fully degraded peptide molecules declines rapidly. Already after seven sequencing cycles this amount is surpassed by the amount of longer (imperfectly degraded) molecules. This complex mixture is reflected in the PTH-analysis chromatograms.

the correctly degraded molecules. The resulting complex mixtures of molecules of different length and terminal amino acid residues, of course, lead in the next degradation cycle to a complex mixture of PTH-amino acids. This leads to an increasingly difficult to interpret PTH analysis. Usually, a sequence is no longer interpretable when the molar amount of the newly degraded amino acid (corresponding to the peptide with the correct length due to 100% complete reactions yield in every step) falls below 15% of the starting amount of peptides. The higher the repetitive yield, the later this value is reached and the longer the sequences that may be obtained (Figure 14.8). Only after the automation of the Edman degradation could repetitive yields over 90% be obtained. Today "normal" yields are about 95%, resulting in average achievable sequence lengths of 30–40 amino acids.

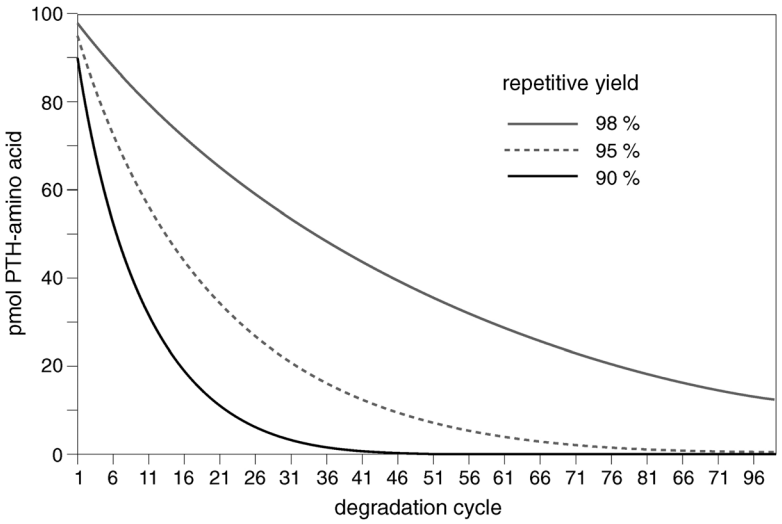


Figure 14.8 Length of amino acid sequence degradation at different repetitive yields. In Edman degradation the number of the fully degraded peptide chain molecules declines rapidly. If the value of the amino acid newly cleaved off drops below 15% of the staring amount, the sequence can usually no longer be interpreted.

Table 14.1 Milestones in amino acid sequence analysis and the amount of material needed for amino acid sequence analysis.

Year	Milestone	Amount needed for analysis
1950	Method of amino acid sequence analysis	100 nmol
1967	Automated sequencer	5 nmol
1971	Solid phase sequencer	
1976	HPLC detection of PTH-amino acids	500 pmol
1978	Polybrene as carrier substance	
1978	Dead-volume-free valve blocks	
1980	Gas phase sequencer	100 pmol
1984	Microbore HPLC detection of PTH amino acids	<10 pmol

14.1.4 Instrumentation

As mentioned earlier, the amino acid sequence analysis as proposed by Edman has remained virtually unchanged since 1950. Pehr Edman published in 1967 an automated version of the method that reduced the losses that inevitably occur during manual handling and hence improved the quality of sequence analysis significantly. The over hundred-thousand-fold increase in sensitivity from about 100 nmol starting material in the publication by Edman in 1950 to about 1 pmol today has essentially been achieved with just a few improvements (Table 14.1): the change of analysis method for PTH-amino acids from thin-layer chromatography to high pressure liquid chromatography, technical improvements to the instrumentation, and the development of gas phase sequencing.

The Edman sequencer consisted of a solvent delivery unit that transported the solvent and reagents with nitrogen pressure via an electronically controllable valve block to a reaction compartment. After the reaction steps of coupling and cleavage, the cleaved ATZ-amino acid was collected in a cooled fraction collector and then fed to the offline conversion and identification steps. This simple arrangement also applies in principle for the modern sequencer (Figure 14.9). In addition, today an automatic conversion unit (introduced in 1976 by Wittmann-Liebold) is used, in which the cleaved ATZ-amino acid is transported to a separate vessel, where the conversion reaction is carried out. Then, PTH amino acid identification is performed online by HPLC.

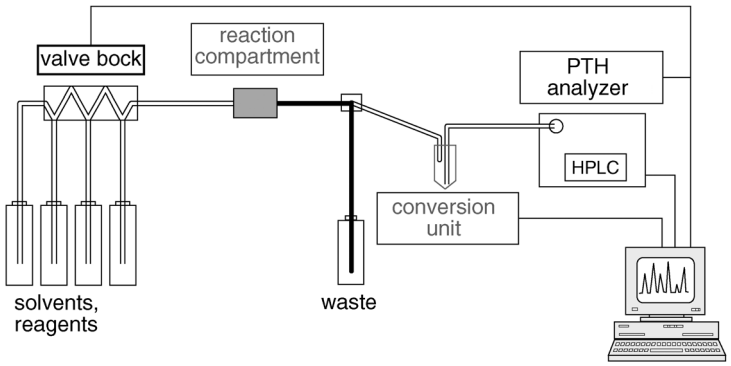


Figure 14.9 Scheme of a protein sequencer. A battery of bottles containing the solvents and reagents are pressurized by argon gas. Delivery is performed by opening of dead-volume-free valves. Defined volumes are delivered to the reaction compartment, where all the reactions of the Edman degradation take place. The released ATZ-amino acid is transported to the conversion unit, converted into the PTH amino acid, and then identified in the PTH analyzer.

The most important part of a sequencer, where the biggest changes have also been made over the years, is the reaction compartment. Its task is, essentially, to immobilize the protein to a defined position so that the various reactions of the Edman degradation can be run reproducibly and controllably. The biggest problem is that proteins are highly soluble in some solvents or reagents of Edman degradation, especially in the base and the acid. Therefore, conditions must be chosen that prevent the washing out of the protein during sequence analysis. Thus, the reaction compartments, depicted in Figure 14.10, determine the different types sequencer.

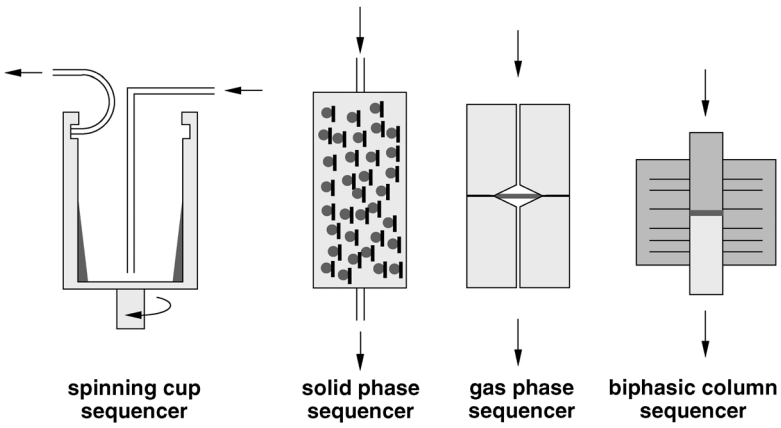


Figure 14.10 Reaction compartments for amino acid sequence analysis of proteins.

Liquid Phase Sequenator Edman used, in the sequenator he presented in 1967, a rotating cup (spinning cup), in which the protein was maintained by centrifugal force against the wall. Base and acid were added in such an amount in the rotating cup that only the protein-containing part was wetted. After completion of the reaction, most of the base or acid was removed via a vacuum system. The extraction of reagents and reaction by-products was carried out very efficiently by continuous addition of the solvent through the central tube and removal by a tube located in a groove on the upper rim of the cup tube. This principle was subsequently adopted in liquid phase sequencers, including a commercial version, but was only used practically until the early 1980s.

Solid Phase Sequencer Especially with shorter proteins and peptides protein wash out during the reactions and washing steps of the Edman degradation is a major concern. Consequently, R. Laursen published in 1971 the principle of solid phase sequencing, in which the problem of the loss of protein due to wash out is avoided by binding the protein covalently to a solid matrix particles. Either reactive side chains of individual amino acids or the free carboxyl group of the C-terminal amino acid are coupled to the solid support (polystyrene, glass) with the help of bifunctional reagents (diisothiocyanate or various carbodiimides), of which one functionality was linked with the matrix (e.g., glass) and the other to the peptide. The solid support with the immobilized protein is then packed into a small column, which is used as the reaction compartment in the Edman degradation.

Crosslinking Reagents, Section 6.3

Advantages:

- Owing to the chemical fixation of the protein to the matrix more drastic chemical conditions for the sequence degradation can be applied without the risk of protein wash out.
- Washing steps can be virtually extended indefinitely and so optimal and reproducible conditions can be achieved.
- High quality repetitive yields of over 96% can usually be obtained.
- Proteins fixed on the C-terminus can be completely sequenced.
- Very long sequences have been obtained (often up to about 80 amino acid residues).

Disadvantages:

- Amino acid residues, with their side chains bound to the matrix, cannot be extracted and therefore are also not detected as PTH-amino acids, which results in a gap in the sequence information.
- The protein must be bound to the matrix by chemical reactions prior to the sequence degradation. The yields of this immobilization may differ greatly and are often unpredictable.

The solid phase sequencer has not found general acceptance because of its serious drawbacks.

Gas Phase Sequencer The biggest advance in sequence analysis was undoubtedly the introduction of gas phase sequencing in 1981 by Hewick and coworkers, a principle that has since replaced the aforementioned forms of sequence analysis because of its simplicity and efficiency. In gas phase sequencing the protein – eventually using a carrier (polybrene) – is applied to a chemically inert glass frit. The two reagents in which the applied protein is soluble, the base and the acid, are conveyed in gaseous form. An argon or nitrogen stream is passed through an aqueous trimethylamine solution or through trifluoroacetic acid and then transported into the reaction chamber. Thus, the desired basic or acidic conditions can be provided in the reaction compartment, without the possibility of the protein being washed out. The reaction by-products and the ATZ amino acid are then extracted in the liquid phase with the organic solvents; however, in these solvents the protein is not soluble. Based on the gas phase principle in conjunction with newly developed dead volume free manifolds a new generation of instruments were produced. These instruments were ideally suited to the needs of ever-smaller amounts of protein. The first instruments of this kind already sequenced amounts of only 100 pmol of protein. After the introduction of further instrumental improvements, such as online HPLC separation and the continually improved HPLC identification of PTH-amino acids, with a gas phase sequencer today, in optimal cases, protein amounts down to 1 pmol can be sequenced.

Pulsed Liquid Sequencer The pulsed liquid sequencer is basically a gas phase sequencer in which a rapid cleavage reaction is achieved by promoting the acid in liquid form. This requires a very precise dosage to wet the protein with the acid, but not to wash it out of the reaction compartment. Optimized washing times, in conjunction with temperature programming for the various reaction stages of the Edman degradation, has allowed the time for one cycle of Edman degradation to be shortened to about 30 min.

Sequencers with a Biphasic Column Reactor Since 1990 a sequencer has been developed and brought to market in which the reaction compartment consists of two chromatographic columns. A reversed-phase column, to which the protein binds very efficiently, when aqueous solvents and reagents are available. The second column consists of silica gel that binds proteins well under organic solvent conditions. This becomes necessary to keep the protein in the reaction compartment when organic solvents for extraction of hydrophobic small reaction by-products or of the ATZ-amino acid have to be used. The protein is introduced at the interface between the two columns; usually, it is applied directly to the reversed phase column, while at the same time salts and polar impurities can be removed by washing with water or 0.1% trifluoroacetic acid. The solvents and reagents can be supplied from both sides of the combined biphasic column in this sequencer-type. Aqueous solutions (base, acid) are always introduced through the silica gel in the direction reversed-phase column and organic solutions on the reversed-phase column in the direction of silica gel, whereby the protein remains well localized, concentrated in the center of the biphasic column. By optimizing reaction conditions yields of over 95% have been obtained.

Despite this good quality, the biphasic concept failed to find a place in a market that was continuously declining due mainly to the increasing competition of peptide structure determination by mass spectrometry.

14.1.5 Problems of Amino Acid Sequence Analysis

Even though the reactions of the Edman degradation proceed in high yield, the reaction conditions are well-studied, and the instruments of the leading manufacturers are consistently mature today, there are in sequence analysis still a number of problems. Especially in the lower picomole range sequencing may sometimes become very difficult and eventually impossible. These problems can be divided into two categories, which are linked on the one hand with the state of the sample and, on the other hand, concern the chemical or instrumental problems of sequence analysis itself.

Problems of the Sample to be Sequenced The sample must be uniform (pure) at the N-terminal. In each cycle of Edman degradation, the N-terminal amino acid derivatives of each protein species present in the sample are cleaved. After extraction and conversion, all these PTH-amino acids, via reversed-phase HPLC, are separated, identified, and quantified. Interpretation of the HPLC chromatograms is not difficult for pure, uniform samples. However, with heterogeneous samples generally several PTH-amino acids in the HPLC chromatograms of each degradation step are detected. For example, if there are two proteins in the mixture, the assignment of PTH-amino acids to the corresponding primary and secondary sequence succeeds only if the amounts of proteins – and therefore the amounts of the corresponding PTH-amino acids – are significantly different. The assignment is already very uncertain when the proteins are present in a ratio of 2 : 1, and practically impossible if the protein levels are even more similar. The main difficulty in interpreting the data from a sequence of mixtures is caused by the different yields of the individual PTH-amino acids. This yield is dependent on the particular amino acid, reaction conditions, and especially the wash-out of the peptide. The wash out losses are mainly dependent on the hydrophobicity and peptide length. Both are normally not known and can be very different for different peptides.

Contamination like salts, detergents, and free amino acids may interfere even when present in small quantities with the chemical reactions of the Edman degradation or prevent the necessary efficient extraction steps.

Contaminations that are mostly polar in character are poorly removed from the reaction compartment by organic solvents, thus affecting the efficiency of degradation over many sequencing cycles. Relatively nonpolar contaminants (e.g., free amino acids of buffers, vessels, etc.) are extracted together with the PTH-amino acids from the reaction compartment and appear in the HPLC chromatogram. They impede or prevent the identification and quantification of individual PTH-amino acids. However, nonpolar contaminants are washed out after a few sequence cycles and then eventually allow a successful sequencing. This is the reason why in publications, even with otherwise good and long sequences, often the first amino acid is not clearly identified.

In the problem area of contamination, solid phase sequence analysis has distinct advantages, since the protein is coupled to a carrier. The immobilized protein can be washed with diverse, and also polar, solvents so that ultimately the sequence analysis can be performed with a pure sample. For the gas phase sequencing special emphasis should be placed on an appropriate sample preparation, and the last step of a protein or peptide purification should be planned so that the sample is preferably salt-free and without contaminations. For peptides, as the last step of purification on a reversed-phase HPLC the use of volatile solvent systems such as 0.1% trifluoroacetic acid/acetonitrile is very suitable. For proteins (especially in very low amounts) that cannot simply be desalted or contaminations removed by chromatography, a simple and good sample preparation method is non-covalent adsorptive immobilization on hydrophobic membranes. This immobilization can also be achieved from polyacrylamide gels by electroblotting of the proteins on a chemically inert membrane (e.g., PVDF). The

proteins are thereby bound so tightly that salts can be washed away easily without the risk of protein loss.

Application of the Sample Application of the sample on the sequencing matrix is a trivial but crucial step for a successful sequence analysis. It is often underestimated that small amounts of protein (microgram, nanogram quantities) can bind very well and rapidly from aqueous solutions to various surfaces (e.g., of glass tubes, Eppendorf vessels, etc.). Often, only by taking special precautions (e.g., dissolving the sample in 95% formic acid, coating of vessels, etc.) can proteins be fully recovered for sequence analysis.

N-Terminal Blocking In the past the biggest problem for amino acid sequence analysis was N-terminal blocking. For the coupling reaction of Edman degradation a free amino group at the N-terminal end of the protein must be present. In all proteins in which this group is modified, a direct sequence degradation is not possible. About 50% of all naturally occurring proteins have such an N-terminal modification (acetylation, formylation, pyroglutamic acid, etc.). Only in the rarest cases can such blocking be chemically or enzymatically removed prior to sequence analysis. Thus, sequence information from blocked proteins is normally only obtained from internal sequences (i.e., by chemical or enzymatic fragmentation of the protein, subsequent separation, and sequence analysis of the resulting fragments). An N-terminal blocking can also be introduced unintentionally during protein purification or sample preparation. The most common causes of such artificial blockages are certain chemicals (e.g., urea at alkaline pH) and impurities in detergents (e.g., in Triton X-100 that that has been stored for a long time), which can react with the N-terminal amino group (e.g., carbamoylation, oxidation).

Quantification A particularly important, and difficult-to-treat, aspect is to quantify the amount of a sample to be sequenced. A good estimate of the amount of protein present is particularly important since, as mentioned above, there is a large number of N-terminally blocked proteins. A "nice" protein band in the gel, a spot in the 2D gel electrophoresis or a symmetrical chromatographic peak per se does not guarantee a uniform substance. At each sequencing it should be checked whether the amount of material used corresponds to the amount of the resulting PTH-amino acid. It should also be borne in mind that usually only about 50% of the protein present is sequenced (initial yield, see below). If in the sequence analysis an unexpectedly small amount of PTH-amino acid is obtained, various causes are possible.

For example, mixture of proteins may be present with one or more N-terminal blocked proteins. Since in principle the amount of a blocked protein cannot be judged from the sequence analysis, the resulting sequence may be derived from a secondary component of the protein mixture. However, since it is usually the main component that is to be analyzed, the risk is that the main component is blocked and the sequence obtained may be derived from a contamination. A precise quantitative assessment for the sequence analysis can assist in this process, to identify such blocked proteins or even to get a hint of the existence of a blocked protein.

Today it can be proven relatively easy if the protein sample corresponds to the amino acid sequence obtained. A sequence analysis of internal peptides obtained after enzymatic cleavage and mass spectrometric analysis of the produced peptides can be carried out with very small amounts of sample and should be performed routinely.

There may be less protein available than expected. In practice this happens quite often, as the common protein determination methods for small quantities either do not function very well or often exhibit very large errors (factor of 10). The only quantitative protein determination method for small amounts of protein is amino acid analysis. However, amino acid analysis is technically difficult to perform, extremely susceptible to contamination, and often consumes a significant portion of the existing material (0.1–0.5 μg); for these reasons it is only rarely applied. In practice, the amount of protein is often estimated from a color in a polyacrylamide gel, but this depends on many factors such as the gel thickness, staining properties of the individual protein, and also on the amount of protein itself, and therefore, great experience and caution are necessary.

Problems of Sequencing

Problem Amino Acids Unfortunately, equal amounts of various amino acids do not appear with the same intensity in the PTH chromatogram. Some amino acids are partially destroyed by the aggressive reaction conditions of the Edman degradation and give several (small) secondary

peaks. For example, up to 80% of serine and up to 50% of threonine are destroyed by dehydration (β-elimination). Underivatized cysteine is almost undetectable as its PTH derivative. From about 30% up to 100% of the tryptophan is destroyed, depending on the sequence position.

Other amino acids are difficult to extract from the reaction compartment because of their high polarity (arginine and histidine). Lysine is particularly sensitive to oxidation and with sub-optimal quality of the solvents can be present in very poor yields. However, with some experience, most of these amino acids can be detected at least qualitatively. For small quantities or with samples that are not entirely pure, considerable problems in creating a safe and unique sequence may arise.

Modified Amino Acids Modified amino acids can be unstable under the drastically alkaline or acidic conditions of the Edman degradation and may appear like a "normal," unmodified amino acid. PTH derivatives of stable modified amino acids can be close to, or at, the chromatographic position of "normal amino acids" and thus lead may to misinterpretations. Since there are only few PTH standard substances of modified amino acids, very few positions of modified amino acids in the PTH-chromatogram are known.

Some modified amino acids like glycosylated or phosphorylated amino acids produce very polar PTH-derivatives, which cannot be extracted from the reaction compartment by organic solvents. Therefore, they cannot be found in the PTH-amino acid chromatogram.

Background The interpretation of PTH chromatograms is complicated by an increasing background later in sequence analysis. This arises because there are more labile peptide bonds (especially aspartyl bonds) in proteins that are hydrolyzed in each degradation cycle, by a few percent, particularly in the cleavage reaction. New peptide fragments are produced and at the resulting free N-termini of these fragments a new sequence degradation takes place, producing PTH-amino acids that appear in the HPLC chromatogram. Typically, during a sequence analysis the PTH amino acid background increases significantly up to approximately the twentieth degradation cycle, passes through a maximum, which often quite well reflects the amino acid composition of the protein, and decreases again towards the end of the sequence analysis, because some of the fragments are by then already sequenced to the end.

Initial Yield For reasons not yet understood, is not the entire amount of protein that is introduced in a sequencer, to be sequenced. The initial efficiency (initial yield) is the amount obtained in the first step of the PTH-amino acid in relation to the used amount of protein. It is normally only about 50%. But this value is dependent on the used protein, the individual sequencer used, and the actual reaction parameters.

Sequencing Parameters The usually freely selectable parameters in a sequencer such as reaction time, amount and flow rates for the individual reagents and solvents, drying time, and temperature have a serious impact on the quality of a sequence degradation, and both the initial yield and the repetitive yield are affected. The parameters are also dependent on the type and amount of the sample and the carrier material used for the sequence analysis. The standard programs supplied with instruments take this into account by using different optimized programs for sequence analysis. Thus, there are various programs for normal applied samples (that is on glass fibers), for samples coated onto PVDF membranes, or for proteins of blotted samples. In addition, there are special programs, for example, for serine- or proline-rich proteins or synthetic peptides that are still coupled to the synthesis resin. As a routine, almost all sequence analyses are carried out with "the" standard program, which is optimized for high repetitive yields with minimum time required.

Purity of Chemicals The chemicals used for sequence analysis must meet extremely high quality standards, since impurities can affect the repetitive yield or lead to interfering peaks in the PTH analysis. Because of the many other possible causes for reduced sequence yield (see above), in practice error diagnosis is extremely difficult and time-consuming. To ensure a consistently high quality of the reagents and solvents they have to be specially purified and subjected to strict quality controls.

Sensitivity of the HPLC System The entire HPLC system for the separation of PTH-amino acids must be routinely operated in a high sensitivity range, which is technically difficult. The commercially available sequencers can sequence protein amounts in the low picomole range. Today, the quantitative determination limit of PTH-amino acids is usually about 1 pmol. Normally, the 20 PTH amino acids (plus some reaction by-products) can be separated in about 15–25 min. The optimized solvent gradient used for this purpose and the temperatures are extremely accurate complied with, since small changes in solvent composition and/or temperature drastically affect the retention behavior of PTH-amino acids. Therefore, a high constancy of solvent delivery and an extremely high reproducibility of the gradient are required.

14.1.6 State of the Art

With the state of the art sequencers currently available on the market it is possible to obtain sequence information from protein or peptide amounts less than 10 pmol. A prerequisite for this is a sample that is salt- and detergent-free, with the N-terminal amino group being free and not inaccessible due to (natural or artificial) blocking for Edman degradation. Moreover, the instruments and chemicals used for sequence analysis must be in optimum condition and are normally used in the lower picomole level. If, for example, in a sequencer nanomole quantities of synthetic peptides are routinely sequenced, the inevitably existing contamination (both in the sequencer and in the PTH-analyzer system) prevents interpretation of the sequence analysis at the lowest picomole level. In ring tests a surprisingly high number of sequence errors are still detected in the sequence analysis of small but sufficient amounts of peptide. These false sequences cause at worst considerable time and financial damage and often require great effort to detect and correct. As a consequence, it is necessary that manufacturers of sequencing instruments find ways to improve the simplicity and robustness of the devices and implement software solutions that help avoid misinterpretations. On the other hand, independent methods such as mass spectrometry and sophisticated protein/peptide separation tools should be available for the operators of sequence core facilities. This may help, in many cases, to detect or solve problem cases. But most importantly, for a protein sequence analysis project to become successful requires a well-considered purification strategy and an optimal sample preparation.

14.2 C-Terminal Sequence Analysis

The characterization of a protein or peptide at its C-terminal end by determining the last amino acid is desirable for many problems. It would be optimal if the polypeptide chain could also be equally effectively sequenced from the C-terminal end. However, none of the currently existing methods for C-terminal sequence analysis gives similarly good results to automatic, stepwise N-terminal Edman degradation in an amino acid sequencer (Section 14.1.3). Advances in mass spectrometry for the structural determination of peptides and the availability of more complete sequence databases mean that C-terminal sequences are now often simply investigated by mass spectrometry alone or by a combination of mass spectrometry and N-terminal sequence analysis. C-terminal sequence analysis currently only has historical significance and is described here only in outline.

14.2.1 Chemical Degradation Methods

Schlack–Kumpf Degradation Chemical C-terminal degradation, which uses a thiocyanate reagent (Figure 14.11), is analogous to Edman degradation, which is carried out with phenyl isothiocyanate. It is based mainly on a method published in 1926 by Schlack and Kumpf. Optimized conditions were developed by Inglis in 1991. Although many chemical methods have been automated, so far, however, the degradation of a polypeptide starting from the C-terminus has never been as successful as for N-terminal sequencing. The main reason is the low reactivity of the C-terminal carboxyl group in contrast to the highly reactive N-terminal amino group in the polypeptides. The carboxyl group must first be activated to enable a

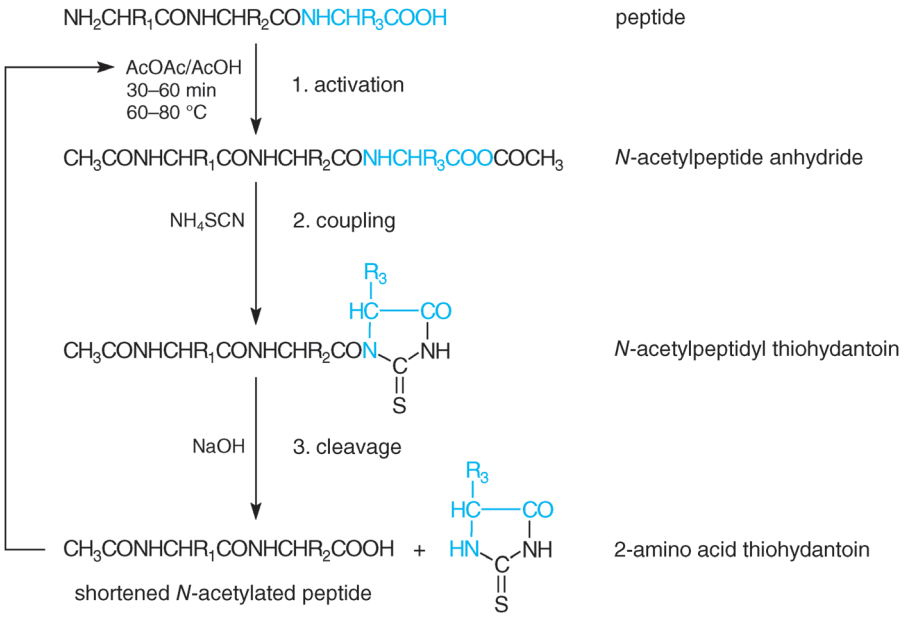


Figure 14.11 Schlack–Kumpf degradation.

coupling with a reagent that is suitable for a gradual degradation from the C-terminal end. For this purpose, relatively drastic experimental conditions are necessary, which are potentially harmful to the partly sensitive amino acids in proteins and peptides. In addition, side chains are modified by the drastic chemical conditions or cleavage of single labile peptide bonds can occur.

C-terminal chemical degradation consists of the following steps (Figure 14.11):

1. activation of the C-terminus,
2. coupling with a thiocyanate reagent to form a peptidyl-thiohydantoin,
3. removal of the C-terminal amino acid as the amino acid thiohydantoin,
4. identification of the cleaved amino acid thiohydantoin (ATH).

Activation A mixed peptide anhydride is produced by acylation with acetic anhydride/acetic acid (AcOAc/AcOH). During this reaction the amino terminal group of the peptide is blocked by acetylation. In addition, internal side chain carboxyl groups may be derivatized depending on the reaction conditions. Peptides N-terminally bound to a solid phase or glass may be derivatized via the side chains of lysine, serine, threonine, aspartic acid, and glutamic acid. The conditions for activation are very drastic: large excess of reagent and reaction times of about 30–60 min at 60–80 °C. Under these conditions, hydrolytic cleavage of acid-labile peptide bonds, for example, cleavage of the Asp-Pro or Tyr-Ser bond, may already occur; consequently, shorter peptide fragments may be formed, which leads to incorrect C-terminal amino acid information.

Coupling Coupling with a thiocyanate or isothiocyanate anion results in a peptidyl-thiohydantoin. For the coupling, various reagents were tested. The most reactive is thiocyanate acid (HSCN). This reagent is very aggressive (highly corrosive to eyes and skin), is unstable, and may not be suitable for use in commercial instruments. Therefore, for the automatic degradation ammonium thiocyanate or guanidine thiocyanate in acetone is used. By addition of acid to these reagents the effective coupling reagent HSCN is generated *in situ* directly in the reaction compartment. Excess reagent and thiocyanate must be removed by argon and with suitable washing steps.

Cleavage The amino acid thiohydantoins are cleaved by treatment with bases or acids. The best method is cleavage with dilute KOH in methanolic-aqueous solution according to Inglis, or in ammonia containing solution, but drastic conditions are necessary (0.1–2 M solutions). A few minutes at room temperature are sufficient for cleavage. Difficulties arise as the rapid temperature change from 60–80 °C to room temperature is technically demanding and the non-volatile KOH

has to be removed by washing steps, whereby the polypeptide, if not covalently bound to carrier, is also washed out.

All steps of the C-terminal degradation method have to be performed under strict exclusion of oxygen. This prevents the oxidation of methionine to methionine sulfone or methionine sulfoxide, the oxidation of tyrosine, the formation of disulfide bridges from cysteines, and the conversion of the thiocyanate reagent into cyanate.

Identification The cleaved amino acid thiohydantoins are separated and identified with reversed phase HPLC on C_{18}-column material (5 µm, 100 A) with a gradient from 0.1% TFA in water/methanol or acetonitrile at 254 nm in a manner similar to that for phenylthiohydantoin (PTH-) amino acids at the N-terminal degradation. In contrast to the phenylthiohydantoins, the ATH-amino acids produced by C-terminal degradation carry no phenyl groups and are significantly more hydrophilic and therefore elute faster and are less well separated.

14.2.2 Peptide Quantities and Quality of the Chemical Degradation

The amount of peptide necessary for a C-terminal sequencer is still within the range of 1–2 nmol, and usually three to five cycles are interpreted. In some cases it was also possible to sequence up to ten amino acids. The difficulties arise from the relatively poor repetitive yield of only 75–85% in the degradation, as losses of peptide occur due to hydrolysis of the labile peptide bonds, wash out, and the destruction of labile peptides and amino acid derivatives. All this results in a high overlap during degradation. Therefore, the degradation of short peptides is more successful than that of the large proteins. In addition, partly uncontrolled degradation, that is, early partial cleavage under acidic conditions followed by a new coupling, results in so-called "prelap" (early appearance of the amino acid of the subsequent degradation step in the preceding step).

14.2.3 Degradation of Polypeptides with Carboxypeptidases

Specificity of Carboxypeptidases To digest the polypeptide from the C-terminal end exopeptidases may also be used. There are various carboxypeptidases, which have different specificities for the individual amino acids: carboxypeptidase A cleaves off at pH 8 neutral amino acids, particularly leucine, phenylalanine, isoleucine, methionine, valine, and alanine; carboxypeptidase B cleaves off only the basic amino acids lysine and arginine, also at pH 8. Carboxypeptidase Y has the broadest specificity and cleaves in both the acidic and the alkaline range (pH 4–8) almost all bonds with the exception of proline. Carboxypeptidase P also cleaves off proline, having a pH optimum at pH 4–5. All carboxypeptidases belong to serine proteases, which indicates that a serine is involved in the catalytic mechanism. Depending on the degree of purity of the enzymes different activities may be observed. The time in which the cleavage of the amino from the C-terminus is achieved varies widely and is concentration and temperature dependent. Therefore, often two to several amino acids are apparently cleaved off simultaneously when one (or a series of) rapidly cleavable amino acid follows one that is cleaved very slowly, such as in the sequence -Gln-Leu at the C-terminal end. Although in such a case treatment with carboxypeptidase cannot deliver the sequence of all amino acids in the digestion, it may give information on which amino acids have been cleaved off. Typically, the C-terminal amino acid sequence with carboxypeptidases is determined in a time experiment, in which amino acids are cleaved off sequentially. For this purpose, the polypeptide is treated with carboxypeptidase for different periods of time. Thus, for example, samples are taken at time 0 (control), 5, 10, 20, 30, 40, 80 min and 2 h and analyzed by amino acid analysis. An example is shown in Figure 14.12. To degrade the polypeptide chain under controlled conditions temperatures of 25–10 °C (or 0 °C) are usually applied, although the optimum for carboxypeptidases is 37 °C.

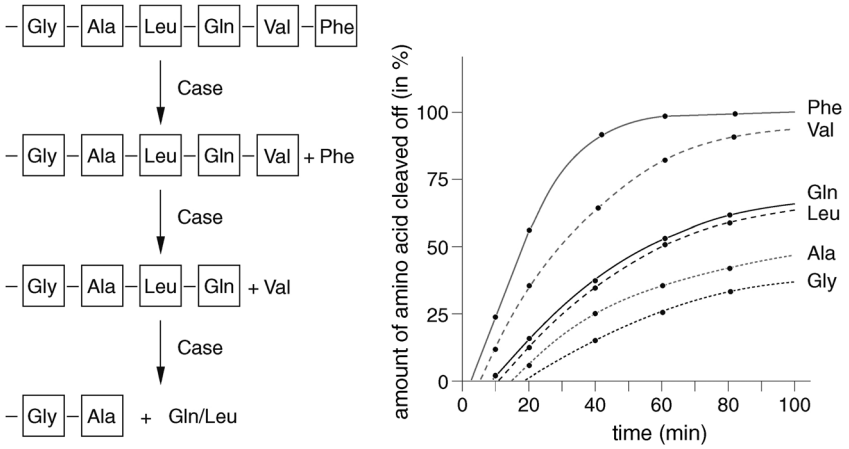


Figure 14.12 Enzymatic degradation of a peptide chain with the C-terminal Gly-Ala-Leu-Gln-Val-Phe by carboxypeptidase Y (Case).

Detection of the Cleaved Amino Acids In the past, the released amino acids were usually determined in the amino acid analyzer. A portion of the sample was brought to pH 2.0 to stop the enzyme activity and was then applied onto the separation column of the amino acid without separation of the residual peptides or of the enzyme. The presence of the proteins does not interfere, as they will stick on the ion exchange column (cation exchanger Dowex 50) used or the reversed phase HPLC (C_{18}, 80–100 Å) and are eluted only during the regeneration of the column. However, the analysis of the released amino acids has two serious drawbacks: First, di- or tripeptides present in the sample are also eluted in the chromatography and can lead to confusion with amino acids eluting in the same place. On the other hand, internal peptide bonds within the protein chain may also be cleaved during the enzymatic digestion. The fragments are also degraded by the carboxypeptidases and thus amino acids from somewhere in the interior are released and detected and distort the results. These internal cleavages have been observed in many proteins, either by contamination of the carboxypeptidases with traces of exopeptidases of broad specificity or due to contamination with traces of endoproteinases. Often, the C-terminal end of the protein is, for steric reasons, not readily accessible to digestion with carboxypeptidases, but nevertheless amino acids can be detected due to internal cleavages. Therefore, the described method using only the analysis of released amino acids is very risky and in principle is no longer recommended. Instead, today it makes more sense to analyze in a carboxypeptidase digestion experiment portions of the complete sample by mass spectrometry. In a simple way we can now determine whether the C-terminal end of the peptide chain has been degraded or whether internal fragments were formed.

Further Reading

Edman, P., and Begg, G. (1967) A protein sequenator. *Eur. J. Biochem.*, **1**, 80–91.

Hewick, R.M., Hunkapiller, M.W., Hood, L.E., and Dreyer, W.J. (1981) A gas-liquid solid phase peptide and protein sequenator. *J. Biol. Chem.*, **256**, 7990–7997.

Inglis, A.S. (1991) Chemical procedures for C-terminal sequencing of peptides and proteins. *Anal. Biochem.*, **195**, 183–196.

Schlack, P. and Kumpf, W. (1926) Über eine neue methode zur ermittlung der konstitution von peptiden. *Z. Physiol. Chem. Hoppe-Seyler*, **154**, 125–170.

Thiede, B., Salnikow, J., and Wittmann-Liebold, B. (1977) C-terminal ladder sequencing by an approach combining chemical degradation with analysis by matrix assisted laser desorption ionization mass spectrometry. *Eur. J. Biochem.*, **244**, 750–754.

Mass Spectrometry

15

Fouzi El Magraoui*, Andreas Schrötter*, Birte Beine, Helmut E. Meyer, and Corinna Henkel

Leibniz-Institut für Analytische Wissenschaften - ISAS - e.V., Bunsen-Kirchoff-Strasse 11, 44139 Dortmund, Germany

Mass spectrometry (MS) is an analysis technique for the detection of *mass to charge ratios* (*m/z*) of ions in a high vacuum. At the end of the 1980s the work of John B. Fenn (USA) led to the development of electrospray ionization (ESI). Franz Hillenkamp, Michael Karas (Germany), and Koichi Tanaka (Japan) discovered the utilization of matrix assisted laser desorption/ionization (MALDI). These new soft ionization methods made it possible to transfer larger molecules like proteins, complex carbohydrates, and long-chain nucleic acids into the gas phase and make them accessible for mass spectrometric analysis. In 2002 Fenn and Tanaka were decorated with the Nobel Prize in Chemistry for their methodological developments in the field of identification of macromolecules. Both ionization techniques are complementary in their requirements and strengths and dominate the current biological mass spectrometry field.

With the help of MALDI- and ESI-MS it is possible to determine the exact mass of biological molecules and thus their chemical composition can be verified or falsified. With the known amino acid sequence of a protein, the difference between a measured and a calculated mass can be indicative for post-translational modifications like phosphorylation or glycosylation. Furthermore, after proteolytic cleavage unknown proteins can be easily and quickly identified. This is based on the measured exact peptide masses and a comparison with sequence databases. Further possibilities arise, for example, out of the combination of two mass analyzers. In this case molecular ions can be analyzed in a first step and selected molecular ions fragmented in a second step (tandem-MS). By doing so it is possible to elucidate the structure of new, so far unknown molecules, or by comparison with databases a quick and safe identification is possible.

For peptide analysis the sensitivity of recent mass spectrometry based approaches is in the sub-femto mole range. For small proteins (<20 kDa) the detection limit is around 10–100 fmol, for mid-sized proteins 50–500 fmol are needed, and 0.5–5 pmol for large proteins (>50 kDa). In general, independent of substance class, the larger the analyte the lower the detection sensitivity. Nevertheless there is a great difference between different substance classes. The mass analysis of nucleic acids and carbohydrates, for example, needs about a 1000-fold higher amount of analytes in comparison to peptides in the same molecular range.

In the simplest case a mass spectrometer consists of an *ion source*, where ions are generated and released, a *mass analyzer*, which separates ions in terms of their mass to charge (*m/z*) ratio, and a *detector* for measuring the ion current (Figure 15.1). As result a mass spectrum is generated, where the relative ion count is plotted against the *m/z* values. In the following section the individual steps of ionization, mass analysis, and ion detection along with the techniques used and functional principles will be explained in more detail.

*Both authors contributed equally to this chapter.

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

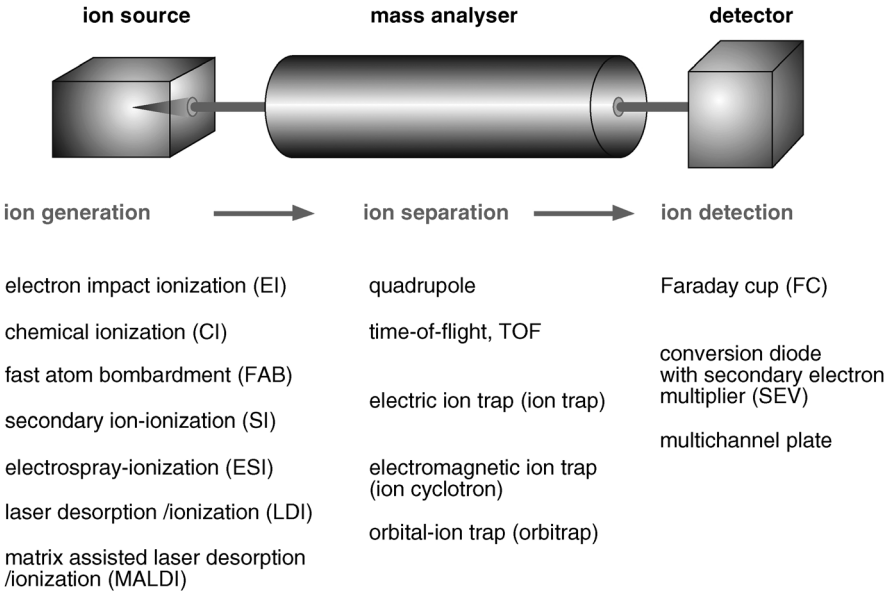


Figure 15.1 Components of a mass spectrometer.

15.1 Ionization Methods

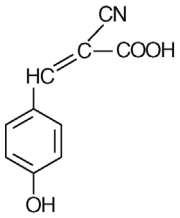
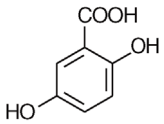
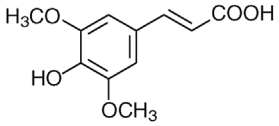
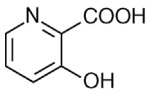
The ionization of analyte molecules in a mass spectrometer can happen more generally through the uptake or loss of an electron, proton, or cation. This can be achieved, for example, by bombardment of the sample by electrons (electron impact = EI), generated reactive ions can transfer their charge to the sample molecule (chemical ionization = CI), or by bombardment with fast atoms (fast atom bombardment, FAB), ions (secondary ion ionization, SI), or photons (laser desorption/ionization, LDI). For polar, non-volatile components and in general for larger biological molecules, ionization of the dissolved sample by ESI or out of a solid phase by MALDI is the method of choice.

15.1.1 Matrix Assisted Laser Desorption Ionization Mass Spectrometry (MALDI-MS)

Since the 1970s lasers have been used in the organic mass spectrometry to reach the goal of direct desorption of intact molecules out of a condensed phase. In these first attempts the sample was prepared as a thin layer on a metal surface and irradiated by a pulsed laser. Normally the generated mass spectra showed weak signal intensity and a pronounced fragmentation of the sample molecule. Since only the detection of ions with a molecular mass of less than 1000 Da was possible laser desorption ionization mass spectrometry (LDI-MS) had no practical impact on the analysis of biomolecules. This changed in 1987 when Michael Karas and Franz Hillenkamp from the University of Münster, Germany, analyzed the interaction between ultraviolet light and organic molecules. They embedded the sample in an appropriate matrix, consisting of small organic molecules, which showed a high absorption at the irradiated laser energy. They observed a significantly better intensity of the analyte ions and nearly no fragmentation in the mass spectra. With this method, named MALDI-MS, it was possible to analyze intact protein molecules for the first time.

If one mixes a sample with a 1000-fold or higher excess of matrix substance on a metal target, where the matrix shows absorbance at the appropriate wavelength (small organic molecules see Table 15.1), a co-crystallization of matrix and analyte on the sample target occurs after evaporation of the solvent. Incorporation of the analyte into the matrix crystal lattice is required for the following MALDI process (Figure 15.2). In the ion source of the mass spectrometer the

Table 15.1 Typical matrix substances for MALDI-MS in biochemical analytics.

Matrix		Wavelength (nm)	suitable for
a-Cyano-4-hydroxycinnamic acid		337, 355	Peptides
2,5-Dihydroxybenzoic acid (DHB)		266 337, 355	Peptides, Proteins, complex carbohydrates
3,5-Dimethoxy-4-hydroxycinnamic acid (sinapinic acid)		266 337, 355	Proteins
3-Hydroxypicolinic acid		337, 355	DNA and RNA

crystalline surface of a prepared sample is exposed to a laser impulse of short wavelength for a few nanoseconds in a high vacuum. The required energy for the ions comes from UV-radiation by resonant excitation of matrix molecules such as the π-electron system of aromatic compounds.

Theoretical calculations suggest that the saved electronic excitation energy within the matrix molecules relaxes into the crystal lattice and causes a strong disturbance of the expansion. Far before the achievement of a thermal equilibrium a phase transition occurs and dissolves the solid surface partly by explosion. Thereby, in addition to matrix molecules sample molecules are released into the gas phase (Figure 15.2). Obviously, the excitation of internal degrees of freedom of the involved molecules is so low that even thermally unstable macromolecules like proteins survive intactly. However, this only applies up to a certain laser power (laser intensity on the sample), within a range of 10^5 to 10^7 W cm^{-2}. By using excessive radiation the sample is destroyed. Numerous experiments suggest that the matrix has an important role in the ionization process of sample molecules. Photoionized, radical matrix molecules result in a high yield of electrically charged sample molecules by proton transfer.

For MALDI-MS impulse a solid state laser in the UV range can be used. For example, a Nd-YAG-laser (yttrium aluminium garnet crystals, neodymium-doped), with impulse times of 5–15 ns and a wavelength of 355 nm (threefold frequency increase) or 266 nm (fourfold frequency increase), or a nitrogen laser with a wavelength of 337 nm and impulse times of 3–5 ns, can be applied. For the infrared range, an Er-YAG-laser (yttrium aluminium-garnet crystals doped with erbium) with an impulse length of 90 ns and a wavelength of 2.94 µm is available. The laser beam is focused by an appropriate optic, as required, to a laser diameter of 10–500 µm on the sample surface. By using a beam attenuator, the laser power on the sample is adjustable in a variable manner. The sample can be controlled by a video camera and the area to be analyzed can be selected. If the sample plate is mounted on a moveable table in the *x*- and *y*-directions, samples can be approached systematically and different positions can be analyzed selectively.

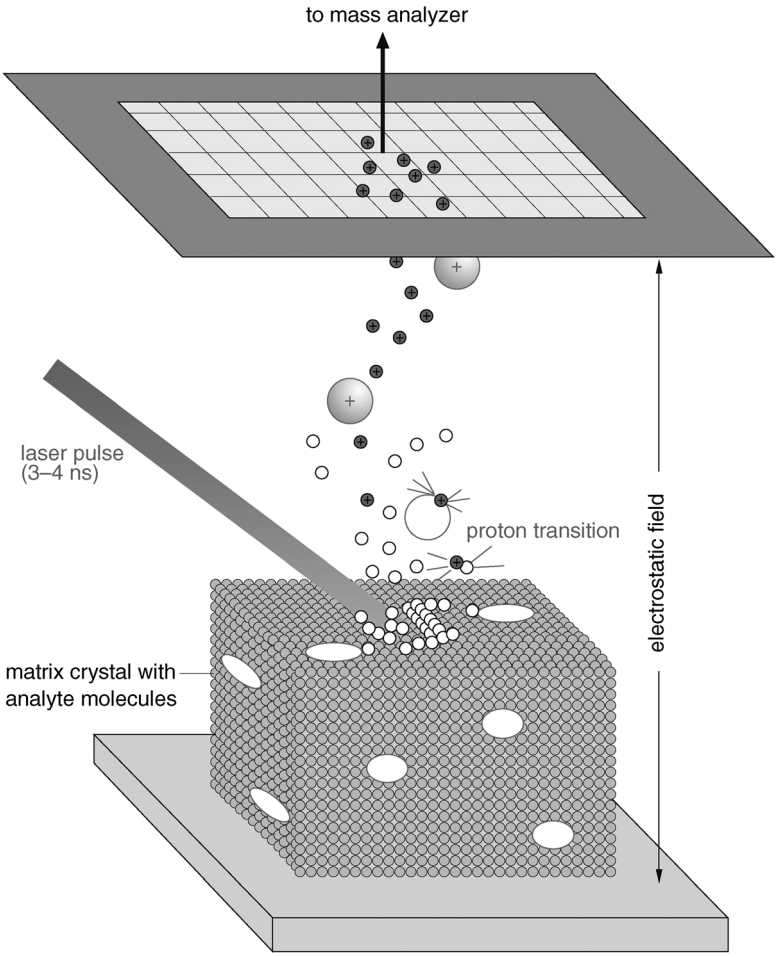


Figure 15.2 Principle of the MALDI process.

Characteristics of MALDI Mass Spectra Figures 15.3 and 15.4 show MALDI spectra acquired in the positive ion mode from a peptide and a protein, respectively, as examples.

For peptides mainly intact single charged ions can be observed. For MALDI analysis of larger molecules, like proteins, there are additional ions beside the single charged, single protonated ones $[M+H]^+$. Additionally, dimer ions $[2M+H]^+$ or multiply charged ions such as $[M+2H]^{2+}$, $[M+3H]^{3+}$, and so on can be detected. The relationship between the appropriate signal intensities is dependent on several factors such as chemical properties of the analyte and its concentration and size. In general, the detection limit is lower with higher m/z values so that the detection of multiple charged molecules is fostered. As a general rule, up to a mass of 50 kDa the highest signal intensity can be observed for the singly charged molecule $[M+H]^+$ in the MALDI spectrum. Beyond that the double charged ion can also dominate the spectrum.

Sample Preparation The principle of MALDI-MS requires the isolation of analyte molecules in a solid, crystalline matrix substance that absorbs the radiated laser photons. The standard MALDI sample preparation is quite easy and requires relatively little effort. One mixes matrix and analyte before transfer to the target, or alternatively directly onto the target (Figure 15.5a below). The sample is then dried, which is why the method is often called the dried droplet method.

Thereby, the matrix is dissolved in pure water or in a mixture of water and an organic solvent (methanol, acetonitrile, etc.) at a typical concentration of about 100 mM of matrix. There is no fixed concentration for analyte solutions, because the specification between different classes of substances can vary greatly. For pure peptide solutions a concentration of 0.1–1 nM is sufficient. In practice higher concentrations are employed or needed (i.e., 10 nM). For proteins

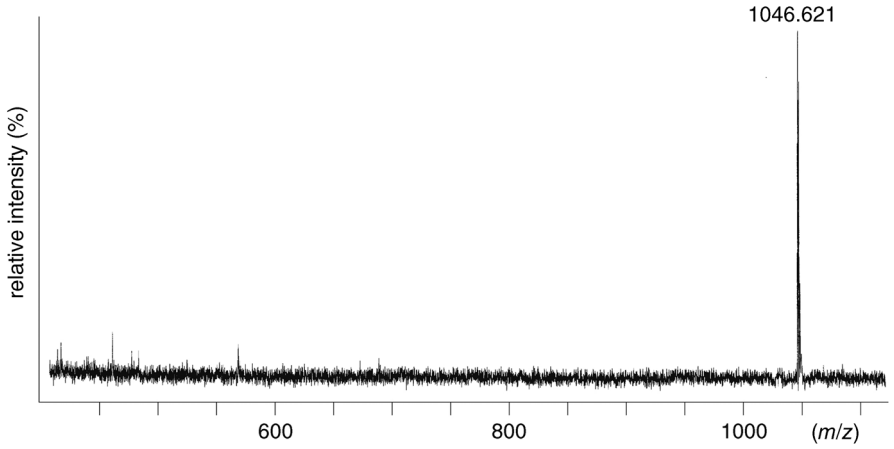


Figure 15.3 MALDI-TOF-spectrum of the peptide angiotensin II.

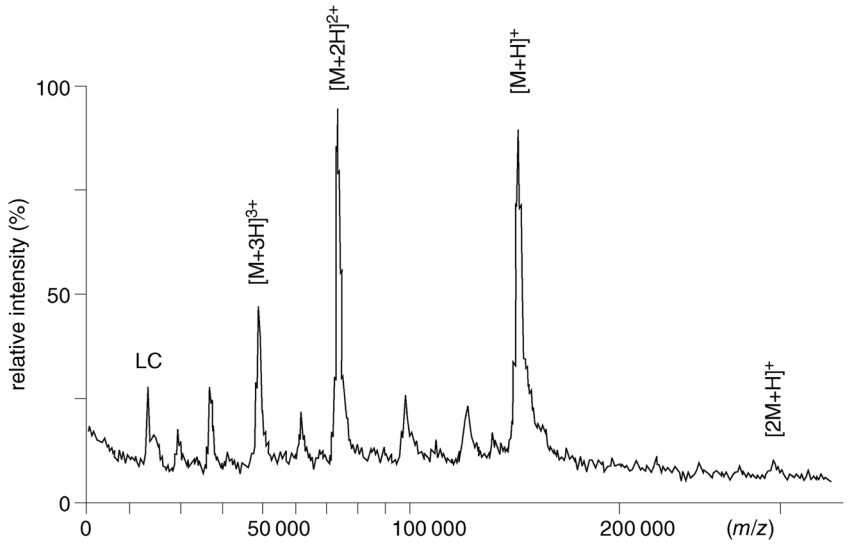


Figure 15.4 MALDI-TOF-spectrum of a monoclonal antibody. Source: By courtesy of K. Strupat and F. Hillenkamp, University Münster, Germany.

the minimal concentrations are much higher, depending on the structure and size (about 0.1–10 µM). For DNA and RNA samples, as for complex carbohydrates, the required minimum sample concentration is set to the same quantity as for proteins. Depending on the analyte the molar ratio of analyte to sample can vary from $1:10^4$ to $1:10^9$.

Sample solutions with too low concentration undercut the ion flow at the detection limit of the detector. If the concentration is too high a high ion flow can cause saturation of the detector, whereby the signal becomes spread and can no longer be evaluated. With higher sample concentration a reduction of ion flow (suppression) is observed. Therefore, it is recommended to test different concentrations of a sample with an unknown protein concentration.

In many cases, especially in MALDI-MS protein analysis, the physicochemical effects during sample preparation are the main constraints on sensitivity. This refers to, among other things, the solubility of peptides and proteins, irreversible adsorption effects, carry-over effects, as well as the influence of contamination of sample containers and solvents. To desalt samples RP-chromatography is often used. Both the preferred ion pair reagent trifluoroacetic acid (TFA) and the solvent acetonitrile are used in the MALDI sample preparation procedure, so that the eluates do not need to be processed further.

Chromatographic Separation Techniques, Chapter 10

In the MALDI-sample preparation a polycrystalline layer of matrix with incorporated analyte arises through evaporation of the solvent. The crystallization process is influenced by the analyte and its physicochemical properties (size, solubility), the choice of matrix and solvent, and the purity of the analyte.

Table 15.1 shows the appropriate matrices for different substance classes (proteins, peptides, lipids, nucleic acids, carbohydrates). Many of the matrices used show crystals 10–100 µm in

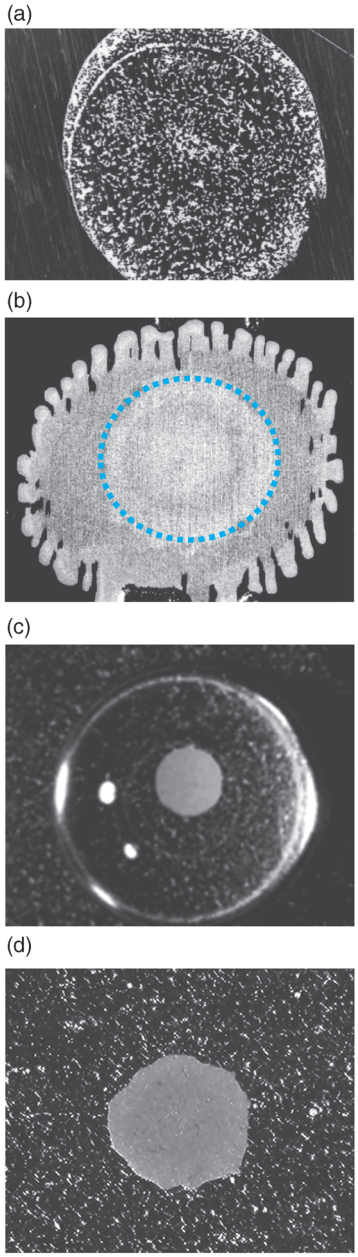


Figure 15.5 MALDI-sample preparation: standard preparation (a), thin layer preparation (b), and thin layer preparation on an anchor chip (c). Source: courtesy of J. Gobom and K. D. Klöppel, Working group mass spectrometry, department Lehrach, Max-Planck-Institute for Molecular Genetic, Berlin, Germany.

size. The crystal lattice is preserved in the typical matrix–analyte ratio ($10^5 : 1$), as could be shown by X-ray structure analysis for some matrices. Nevertheless, the mixing ratios of both components can be different, so that a more or less pronounced local dependence of the analyte integration can occur.

By using other preparation methods such as, for example, thin layer preparation one tries to reach a homogenous matrix crystal distribution on the target. Very thin matrix layers occur, if a matrix that is insoluble in water is chosen, such as, for example, α-cyano-4-hydroxycinnamic acid or sinapinic acid. These are dissolved in a volatile organic solvent such as acetone or tetrahydrofuran and an aliquot of the mixture is applied on the sample carrier. Owing to the low surface tension of the solvent the solution spreads rapidly on the sample carrier and, instead of a mounted droplet, a thin fluid layer is formed. The additional solvent evaporates quickly, mostly in a few seconds, with the consequence that the solution is saturated and starts to crystallize. As a result a homogenous microcrystalline matrix layer develops instead of the deposition of single oversized matrix crystals such as occur in the standard matrix preparation. Thereafter the analyte layer is applied on the matrix layer (Figure 15.5b). As a result the analytes becomes incorporated in the upper layers of the matrix crystal lattice and, for example, bind there. Salts and other strongly polar substances, for example, buffer components, stay as a residue of crust of salt. These residues can be eliminated effectively by washing using acidic water (0.1% TFA-solution). For this purpose the sample is covered for 1–3 s with 3–5 µl of the washing solution and the supernatant is removed afterwards by using a pipette. The washing solution should not contain a noticeable amount of solvent and the pH should be below 2.5, otherwise the matrix crystals can be redissolved. Even if a very small amount of matrix crystals is redissolved, the analyte molecules can be lost completely. The dependence on pH is dependent on the matrices used; whereas the acid is not, or hardly, soluble in water the anion is readily soluble.

A thin layer preparation generally provides a homogenous ion current over the whole sample and therefore facilitates automatic sample measurement. For specific analytes the detection sensitivity and mass resolution can be improved by this kind of technique in comparison to standard preparation methods. This is particularly applicable to peptides. The higher detection sensitivity is based on the fact that the analyte molecules do not become spread over the whole crystal volume but are only incorporated close to the surface. Higher mass accuracy results from the flat elevation profile, which means the time-of-flight differences are lower.

Interestingly, the above-mentioned advantages do not count for intact proteins. They only produce efficient results if the solutions are ultrapure, which means no salts, detergents, or other additives. The standard preparation method is less sensitive to contaminations. But even if the purity criteria are met, the detection limit decreases with increasing protein size so that the detection limit for proteins >20 kDa is much lower compared to standard preparation methods. A remaining benefit of using thin layer preparation is the homogeneity and the associated low dependence on ion current from laser irradiance.

The quality of the sample carrier surface has a direct influence on sample preparation. It influences the spot size of the prepared sample, the crystallization of the matrix, and additionally the possible source of contaminants and carry-over effects if the sample carrier is used several times. They are in general rectangular, planar plates generated out of high quality stainless steel or aluminum with a nickel-plated surface. The surface must be chemically inert, should be easy to clean and typically is divided into a specific raster with marked points for the ordering and assignment of specific sample positions. Typical rasters are 10×10 (100 samples), 8×12 (96 samples, distance: 9 mm), 16×24 (384 samples, distance: 4.5 mm), and 32×48 (1536 samples, distance: 2.25 mm). The last three correspond to the specification of standard microtiter plates (MTP format) and simplify the usage of pipetting robots and in general the organization of sample preparation in high throughput experiments. In principle, current MALDI TOF mass spectrometers are able to process more than 1000 samples in a fully automated way per day. In practice this is still an exception; the analysis of several hundred samples has become more common.

To simplify the sample throughput and to enhance the sensitivity, special sample carriers were developed; the surface of the carrier can influence the sample preparation in a positive way. The surface has a strongly hydrophobic, several micrometers thick, layer similar to Teflon. This is interrupted by defined circled hydrophilic spots with a diameter of 200, 400, 600, or 800 µm. These are used as sample anchor and the prestructured sample carriers are named

anchor chip (AnchorChip™) targets and can be combined with both of the above preparation techniques. This allows us to predefine the sample spot size in advance. This results in an exact sample raster and size, which enables automatic measurement and lead to shorter measurement times. Additionally, the defined spot size allows the deposition of bigger volumes of sample and matrix on the same spot size and therefore leads to better detection sensitivity. Figure 15.5 illustrates the underlying principle.

To enhance the detection sensitivity of standard preparations of peptides, proteins, or nucleic acids with hydrophilic matrices, such as DHB or 3-hydroxypicolinic acid, the concentration of analyte and matrix has to be reduced. Consequently, the sample and matrix are concentrated on the sample carrier before crystallization. This procedure requires contaminant-free samples, because the concentration of contaminants will also increase. For insoluble matrices, such as α-cyano-4-hydroxycinnamic acid or sinapinic acid, a prestructured sample carrier in combination with thin layer matrix preparation leads to the highest detection sensitivity. This combination is mainly used in peptide analyses.

15.1.2 Electrospray Ionization (ESI)

The term electrospray describes the dispersion of a liquid in a large number of small charged droplets in an electrostatic field. This phenomenon had already been observed in the last century and became the basis for many technical applications such as the painting of surfaces. At the beginning of the 1970s Malcolm Dole and his staff used the electrospray method to measure molecular masses. In these experiments polystyrene oligomers 50–500 000 Da in size were sprayed via a needle in a nitrogen-filled chamber using volatile solvents. To generate the spray a potential of several thousand volts between the needle tip and the chamber was required. Dole realized that the charge density on the surface increased with increasing evaporation of the solvent and that this had to be the reason for the explosive decomposition of a drop. He further argued that – at sufficient dilution of the sample – an extremely small droplet with only one macromolecule ion undergoes a series of successive decays.

The remaining solvent molecules would further evaporate while a portion of the charges would remain on the macromolecule. Owing to the lack of suitable mass analyzers Dole was not able to directly detect these large ionized macromolecules. A decade later the research group of M. Alexandrov at the University of Leningrad and the group around J. Fenn at Yale University resumed investigations of the electrospray process by using small molecules. To analyze ions they used a so-called quadrupole mass spectrometer, which enabled them to understand and optimize the electrospray process in more detail. In the mid-1980s both groups were able to show that the electrospray process resulted in a defined ionization and complete desolvation of analyte molecules sprayed in solution. This work led to the establishment of electrospray ionization mass spectrometry (ESI-MS).

Principle of Ionization Desolvation, the transfer of ions from the solution into the gas phase, is an endergonic process. For example: the free energy is needed to convert one mole of sodium ions from aqueous solution into the gas phase is very high:

Endergonic: works only with the supply of energy

$$Na^+(aq) \rightarrow Na^+(g); \quad \Delta G^\circ_{sol}(Na^+) = 410.9\,kJ\,mol^{-1} \tag{15.1}$$

Ionization methods such as LD (laser desorption), FAB (fast atom bombardment), and PD (plasma desorption) are the source of the energy required for the transfer of the molecules into the gas phase and their ionization by complex cascades of high energy collisions and local deposition of energy.

In contrast, ESI leads to desolvation of the dissolved ions. In this mode only a small part of the energy is transferred to the ions. In the electric field the ions (Figure 15.6) are transferred at atmospheric pressure in the gas phase. This process as described by Dole is divided into four distinct steps:

- the formation of small charged droplets out of the electrolyte solution;
- continuous solvent loss by evaporation of the droplets, which leads to an increase in charge density on the droplet surface;

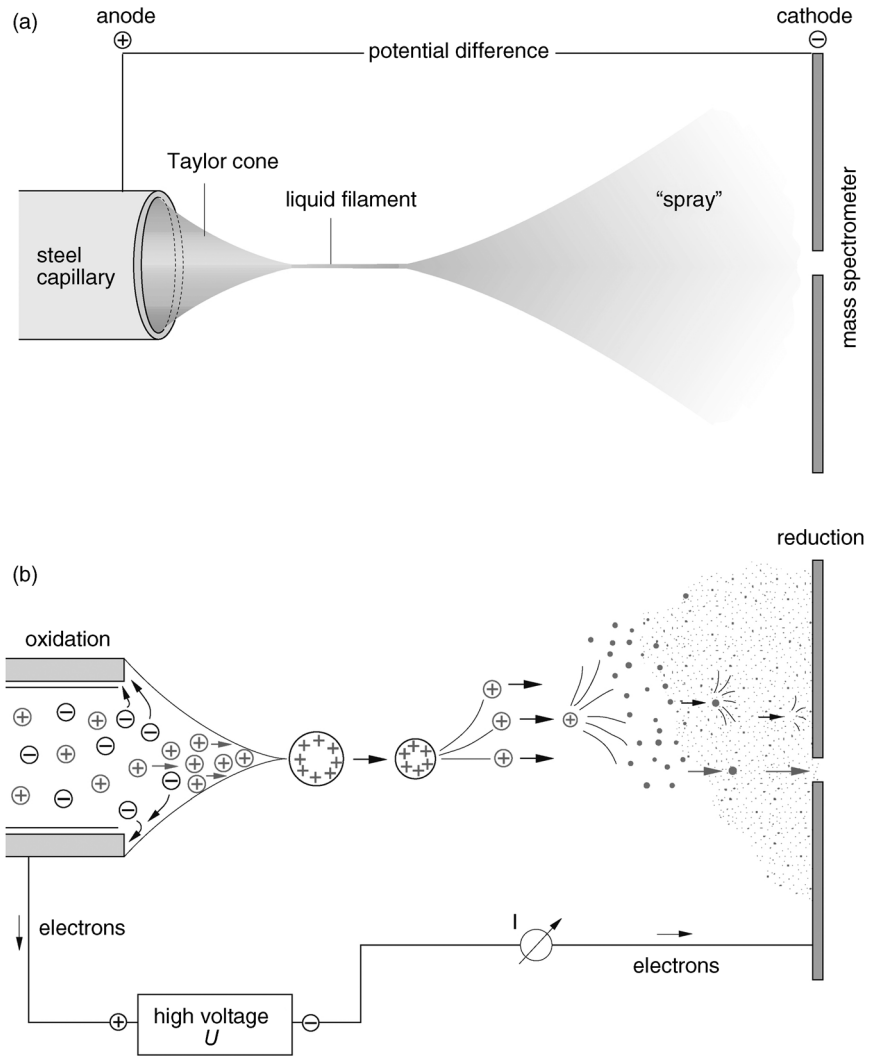


Figure 15.6 Schematic representation of the macroscopic (a) and microscopic ESI process (b).

- repeated spontaneous disintegration of the droplets into micro-droplets (coulomb explosion);
- finally, desolvation of the analyte molecules during the transfer into the mass spectrometer.

As shown in Figure 15.6, schematically for the detection of positively charged ions, the ESI process begins with the continuous supply of dissolved analytes at the tip of a conductive capillary. The electric field applied between the capillary and the mass spectrometer penetrates the analyte and separates the ions, similar to electrophoresis. The positive ions are attracted to the liquid surface. Accordingly, the negative ions move in the opposite direction, until the electric field in the liquid is neutralized by the redistribution of negative and positive ions. Thereby, other possible forms of ionization can be suppressed, such as ionization by removing an electron from the analyte (field ionization) at very high electric fields.

The surface accumulated positive ions of the liquid are drawn further towards the cathode. This results in a characteristic cone liquid (Taylor cone), which is formed because of the surface tension of the liquid, which counteracts the electric field.

At a sufficiently high electric field the cone is stable and emits from its tip a continuous filamentary flow of liquid a few micrometers in diameter. This flow is unstable at some distance further from the anode and is divided into tiny concatenated droplets. The surface of the droplet is enriched with positive charges. Since there is no negative counter-ion there is a net positive charge. The electrophoretic separation of the ions is responsible for the charge in the droplets. The positive ions (as well as, after reversal of the field, negative ions) observed in the mass spectrum are always the ions already present in the (electrolyte) solution.

Additional ions are observed only at very high voltages as a result of electrical discharges on the capillary (corona discharges). The charge balance in the ion source results from the chemical oxidation at the positive electrode and the reduction at the negative electrode.

Experimental investigations have shown that the initially formed droplets have a diameter of a few micrometers and a high charge density ($\approx 10^5$ charges per droplet). These droplets in terms of their composition, size, and charge are close to the stability limit (Rayleigh limit). This stability limit is determined by the repulsive coulomb force of like charges and the cohesive surface tension of the solvent. The Rayleigh equation indicates when the charge Q balances the surface tension γ:

$$Q^2 = 64\pi^2 \varepsilon_0 \gamma r^3 \tag{15.2}$$

where ε_0 is the dielectric constant of a vacuum and r is the radius of the droplets. The droplets shrink by evaporation of the solvent at constant charge Q until the radius r exceeds the Rayleigh limit. After this, they break through the repulsion of like charges into many small droplets of only a few nanometers in diameter (coulomb explosion).

There are currently two model concepts that describe the final formation process of free gas-phase ions.

The older one originated from Dole and is called the "model of the charged residue (charged-residue model, CRM)". This was adapted and further developed by Friedrich Roellgen, resulting in the (single ion in droplet theory, SIDT). The main idea of this model is that the initial droplets undergo a series of consecutive coulomb explosions that result in the formation of very small droplets of about one nanometer radius, which contain only a single analyte molecule. Free gaseous ions are formed by desolvation as a result of collisions with the nitrogen molecules of the gas stream (curtain gas) at the interface to the mass spectrometer (see below). The other mechanism was proposed by J. Iribane and B. Thomson and is called the ion emission model (ion evaporation model, IEM). The central idea of this theory is the direct ion emission from highly charged droplets that contain many analyte molecules. Such droplets have a radius of about 8 nm and contain about 70 charges. Under these conditions (above the Rayleigh limit) free ions are emitted into the gas phase. Despite the decrease in the charges the ion emission is maintained by the continuous decrease of the droplet radius due to evaporation of the solvent. Thus, the IEM comes without the restrictive assumption of only one analyte from very small droplets. There are several experimental observations that can be easily explained by the one or the other model. For example, the occurrence of low charged protein species and adducts of intact protein molecules can be explained by the CRM model. The observation of non-covalent interactions between macromolecules in ESI-MS explains the loss of solvent molecules from nanodroplets better than the ion emission model. Other ESI phenomena can be better explained by the IEM theory.

Figure 15.7 illustrates, as the cause of the coulomb repulsion, the fixed equidistant position of the excess charges on the surface of the droplets. An analyte that reaches the surface of a droplet can take, depending on its spatial extent, positive charges in its emission. A larger spatial extent should therefore result in a greater number of transferred charges. This can be experimentally observed in disulfide bonds cleaved or denatured peptides that show a shift towards higher charge states.

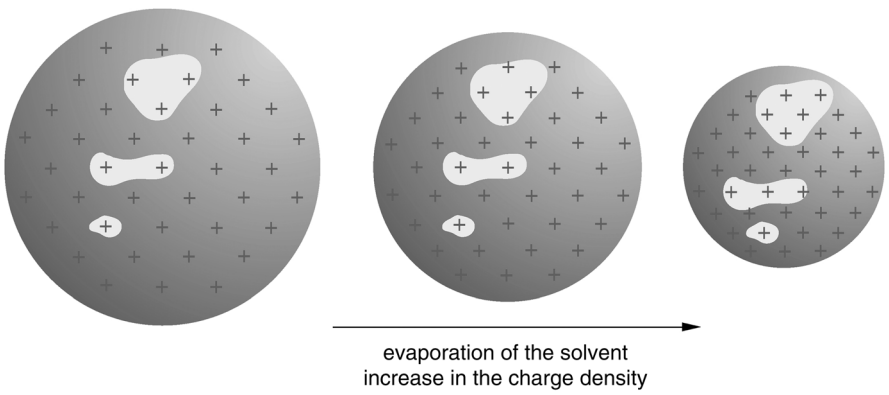


Figure 15.7 Ion emission model (IEM) in the ESI. The droplets produced by electrospray have excess charges, which occupy an equidistant location on the surface due to the coulomb repulsion. By evaporation of the solvent the charge density on the droplet surface increases. Prior to ion emission, the analyte molecules are loaded with the maximum number of charges that they can be held according to their spatial dimensions. Source: Fenn, J. *et al.* (1993) *J. Am Soc. Mass Spectrom.*, **7**, 524–535. With permission © American Society for Mass Spectrometry 1993.

The charge distribution and the fact that small molecules preferably carry only a few charges in contrast to large molecules that carry more charges can be explained by the IEM theory as follows: Owing to the loss of solvent molecules the droplet radii shrinks while the surface density of charges increases (Figure 15.7). For the ion emission of a molecule out of a droplet with a relatively large radius, the number of transferred charges is smaller than the emission from a droplet having a relatively small radius. The presence of a continuous distribution of the droplet radii, which contribute to a certain distribution of ion emission, results in the observed distribution of charge states.

Estimations indicate an electrospray ionization probability between 0.01 and 0.1. This results in a relatively efficient ion formation of the dissolved analyte molecules. During the desolvation process the evaporating solvent molecules lose thermal energy, which results in the formation of so-called, energetically, relatively "cold" ions. This is why ESI is a much gentler ionization method than, for example, MALDI.

ESI-Source and Interface ESI takes place at atmospheric pressure; in contrast, the subsequent analysis of the free ion occurs in a high vacuum ($\leq 10^{-5}$ Torr). This requires a special interface to ensure the transfer of ions into the mass analyzer. Figure 15.8 shows schematically the structure of an ESI source with an interface to a quadrupole mass analyzer.

The ionization chamber and the mass spectrometer are connected through a pinhole (diameter 100 µm). Dry, heated nitrogen (e.g., 60 °C) flows between the outlet and the much larger opening of the interface plate in the ionization chamber.

Nitrogen (also referred to as the curtain gas) collides with the molecular complexes of the electrospray. This prevents a large number of neutrals being drawn into the high vacuum. The collisions also support the desolvation of the ions. Instead of a simple aperture, some devices use an approximately 20 cm long heated curtain gas (≈200 °C) with a diameter of 400 µm. The microdroplets of electrospray can be desolvated effectively by passing them through the transfer capillary.

The electrospray source is basically a capillary through which the analyte is continuously injected into the electric field. This capillary simultaneously forms the interface to the counter electrode plate to which the potential difference is generated in the ionization process. Figure 15.9 illustrates common variants of electrospray sources. For the use of higher flow rates (e.g., as in direct coupling with the LC) a pneumatically assisted electrospray source has been developed. For that an additional coaxial steel capillary leads a gas stream of nitrogen or synthetic air to the top of the source to atomize efficiently the analyte solution at the outlet. Such

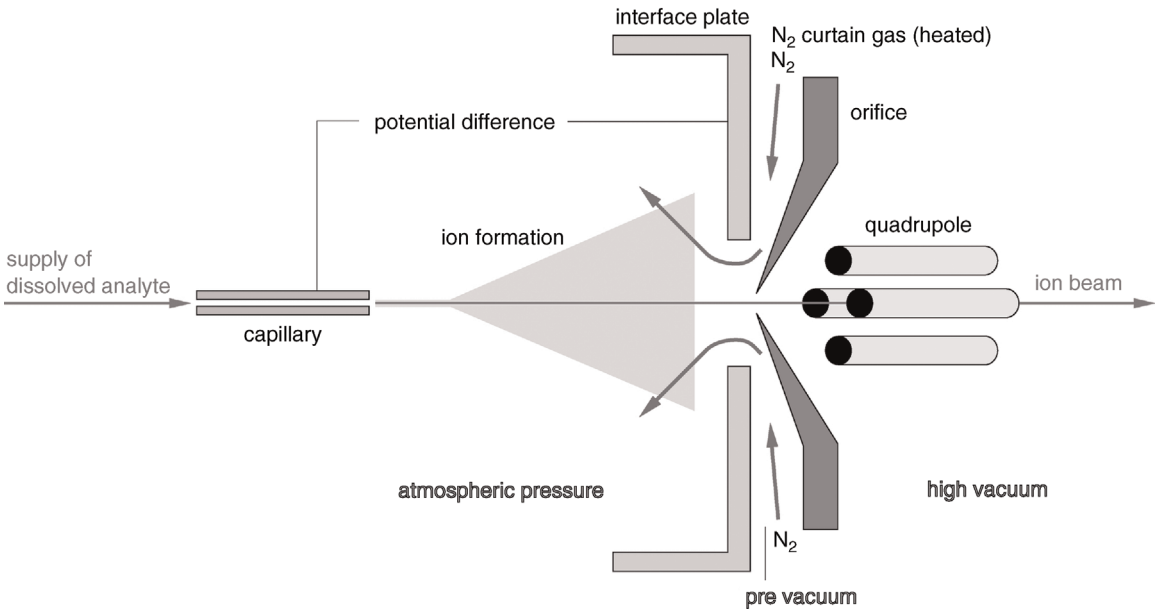


Figure 15.8 Construction of an ESI source with an interface to a quadrupole mass spectrometer.

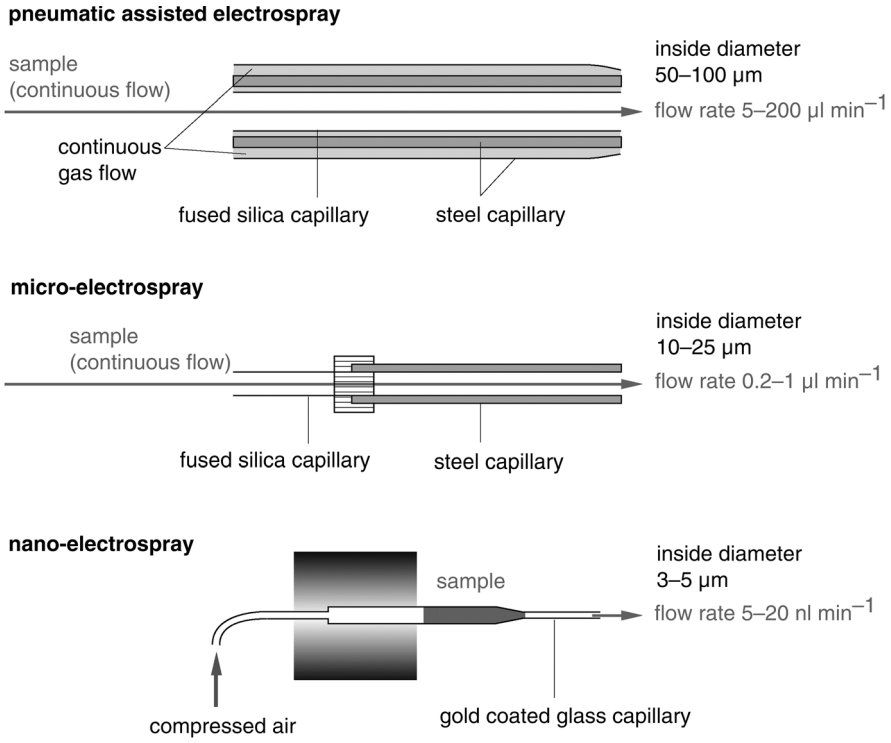


Figure 15.9 Schematic structure and variants of electrospray sources.

sources are used in many devices as standard and enable flow rates from 5 µl up to 1 ml min^{-1}. In the literature pneumatically assisted ESI sources are often referred to as ion spray sources.

Sensitivity studies showed that the observed ion current is correlated with the analyte concentration rather than the unit of sprayed solution. At a high flow rate of 1 ml min^{-1} the same ion intensity is obtained as a flow rate of 5 µl min^{-1}. This means that with increasing flow rates an increasing proportion of the spray in the ionization chamber and on the interface board disappears. The characteristic of the electrospray process suggest that the reduction of the flow rate results in an increased sensitivity of the mass spectrometric analysis.

This has triggered the development of micro- and nano-electrospray sources with much smaller flow rates (Figure 15.9). They have become especially important in the online coupling of ESI-MS and nano-HPLC systems with flow rates around 200 nl min^{-1}. Thus, the outlet opening of the tapered capillary is only a few micrometers in diameter in the nano-ESI-source. In this case, 0.5–3 µl of the analyte solution can be injected directly into the ESI capillary for the analysis of individual samples. In this sample application mode (also referred to as offline nano-ESI) a low gas pressure promotes the sample continuously at a few nl min^{-1} in the electrical field, wherein very small primary droplets with a diameter of ≈200 nm are generated. Owing to this very low flow rates a corresponding long acquisition time for even small amounts of sample can be obtained.

This means that the sample analysis of a solution of 1 µl (typical concentration of 1–100 fmol $µl^{-1}$) at a flow rate of 50 nl min^{-1} can be measured for a maximum time of 20 min. This can be very advantageous for optimizing the measurement conditions and the accumulation of spectra of weak intensity (e.g., MS/MS spectra). A disadvantage of fine capillary tips is the risk of clogging by microparticles or crystallization of the analyte.

Properties of ESI Mass Spectra The formation of multiply charged ions is characteristic for the ESI process. In the mass spectrum every peptide and protein, according to their molecular weight, results in a series of ions with different charges. This occurs, usually, by the addition of a proton in the positive mode or in the negative mode by the subtraction of a proton.

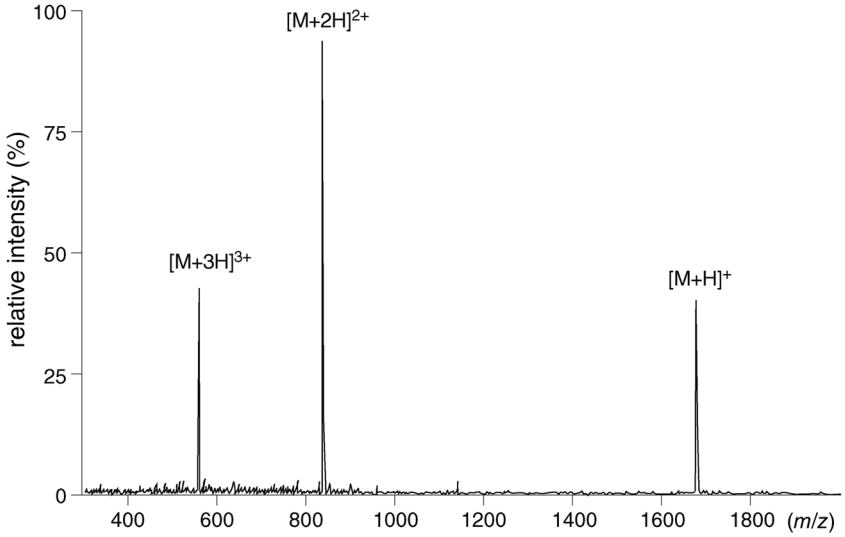


Figure 15.10 ESI-spectrum of the peptide neurotensin.

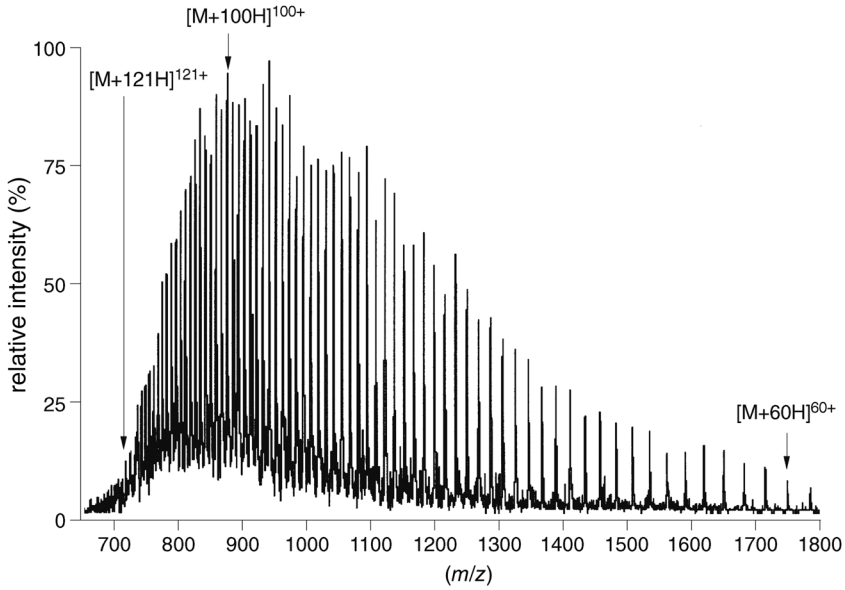


Figure 15.11 ESI-spectrum of the protease protein LA from *Escherichia coli*.

Figure 15.10 shows the ESI-MS spectrum of neurotensin as an example of a peptide; Figure 15.11 shows the spectrum of a protein (protease LA).

The charge distribution of the molecular ions depends on several factors:

With increasing mass and number of basic or acidic functional groups the average number of charges increases. This effect also depends on the choice and composition of the solvent. Using a RP-HPLC in combination with an acidic buffer system it can be observed that peptides with a molecular weight below 1000 Da are predominantly singly charged, while those above 2000 Da are predominantly doubly charged ions. In the example shown in Figure 15.10 the doubly protonated ion at *m/z* 836.9 is the dominant ion followed by additionally two signals for the singly and triply charged ion at *m/z* 1672.9 and 558.3.

The spectra of proteins exhibit a characteristic, approximately bell-shaped charge distribution of the molecular ions.

Cleavage of Disulfide Bridges, Section 9.4

The maximum of the distribution is a function of the parameters of the ESI-mass spectrometer (aperture voltage, density of the gas curtain, etc.), the pH of the solvent, and the denaturation state of the protein.

After cleavage of disulfide bridges and denaturing their structure, the structure of individual proteins becomes more extended. Consequently, charges can be added to or eliminated from free functional groups. The maximum of the charge distribution is shifted to higher charges.

In addition, several factors can have a negative impact on the electrospray process, such as the buffers, salts, and detergents used. Even strong ion signals are often an ion series that is based on the formation of clusters or aggregates (dimers, trimers, etc.) High sensitivity and significant spectra are usually only obtained when the analyte solutions are largely free of buffers, salts, and detergents. Therefore, some samples, like biochemical ones, need to be purified by various approaches.

Sample Preparation According to the principle of ESI the analyte molecules must be dissolved to be transported to the tip of the spray capillary.

Chromatographic Separation Techniques, Chapter 10

Typical solutions are usually mixtures of dipolar organic solvents (methanol, ethanol, acetonitrile, etc.). For the detection of positively charged analyte (positive mode), dilute aqueous acids are used (e.g., 0.01–0.1% formic acid, acetic acid, etc.). The high vapor pressure of the organic solvent, the evaporation process, and the acidic environment supports the protonation of the analyte in the ESI process. Additives such as buffers, salts, and detergents interfere with the electrospray process. The ionization efficiency is already decreased by low concentrations of buffers and salts (>0.1 mM) or detergents (>10 µM). The common concentrations that are used in biochemistry buffer systems (e.g., 100 mM phosphate buffer, 150 mM sodium chloride, 100 mM TRIS buffer, etc.) and the presence of detergents interferes with the formation of free analytes.

Considering the conceptual model for the ionization process we can explain this interference as follows. Ion emissions from the salt and from the buffer ions can compete with the ion emission of the analyte ions. They also tend to crystallize around the analyte molecules. Detergents that are used for the solubilization of proteins can interfere with the formation of free gaseous analyte due to their strong intermolecular interaction. Their activities at the surface of the microdroplets reduce the ion emission. The best and most important method for the preparation of biological samples for ESI-MS is RP-nano-HPLC. The buffer systems and the eluent used have to be compatible with the electrospray process. The often used ion pair reagent TFA has a negative influence on the ESI process and it is therefore only used in very low concentrations (e.g., 0.05%) or is replaced by formic acid or acetic acid.

15.2 Mass Analyzer

There is a huge variety of mass analyzers available for ion detection in mass spectrometry, which all differ in their physical principle for mass analysis. The separation of ions in respect to *m/z* can be accomplished using:

- the ion time-of-flight detector after pulsed ion generation (time-of-flight analyzer);
- the radiofrequency field of a quadrupole system (quadrupole analyzer);
- an ion trap; ion traps can be divided into magnetic-, electric-, and orbital-ion traps; detection is classified as due to either ejection of the ions out of the trap or the ions characteristic resonance frequency.

It is possible to combine different mass analyzers in one hybrid instrument. Such combinations are often used in the field of biological mass spectrometry. All mass analyzers can be combined with either MALDI or ESI ion sources. By coupling these systems it is important to note that MALDI generates ion pulses whereas ESI produces a continuous ion current. Based on the physical differences of the ion sources the mass analyzer requirements are different, too. TOF analyzers require timely defined ion pulses and were therefore originally combined with MALDI. Quadrupole instruments need a continuous ion current and hence are used with ESI ion sources. These limits have been overcome. Using MALDI with high frequencies (e.g., 1 kHz) enables an ion cloud connection leading to a continuous ion current (Figure 15.12a). Equally, a continuous ion current can be analyzed via TOF by separating the ion current using a number of orthogonal orientated electric fields (Figure 15.12b). For analysis of the separated ion current a TOF analyzer with a frequency of, for example, 1 kHz is required.

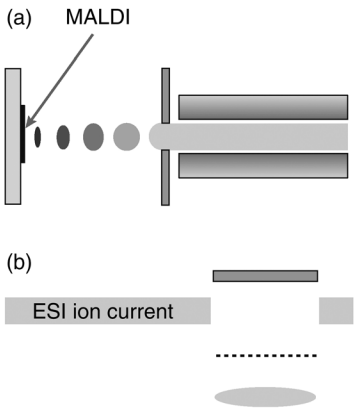


Figure 15.12 Using MALDI with high frequencies (e.g., 1 kHz) enables an ion cloud connection, which leads to a continuous ion current (a). Equally, a continuous ion current can be analyzed via TOF by separating the ion current using a number of orthogonal orientated electrical fields (b).

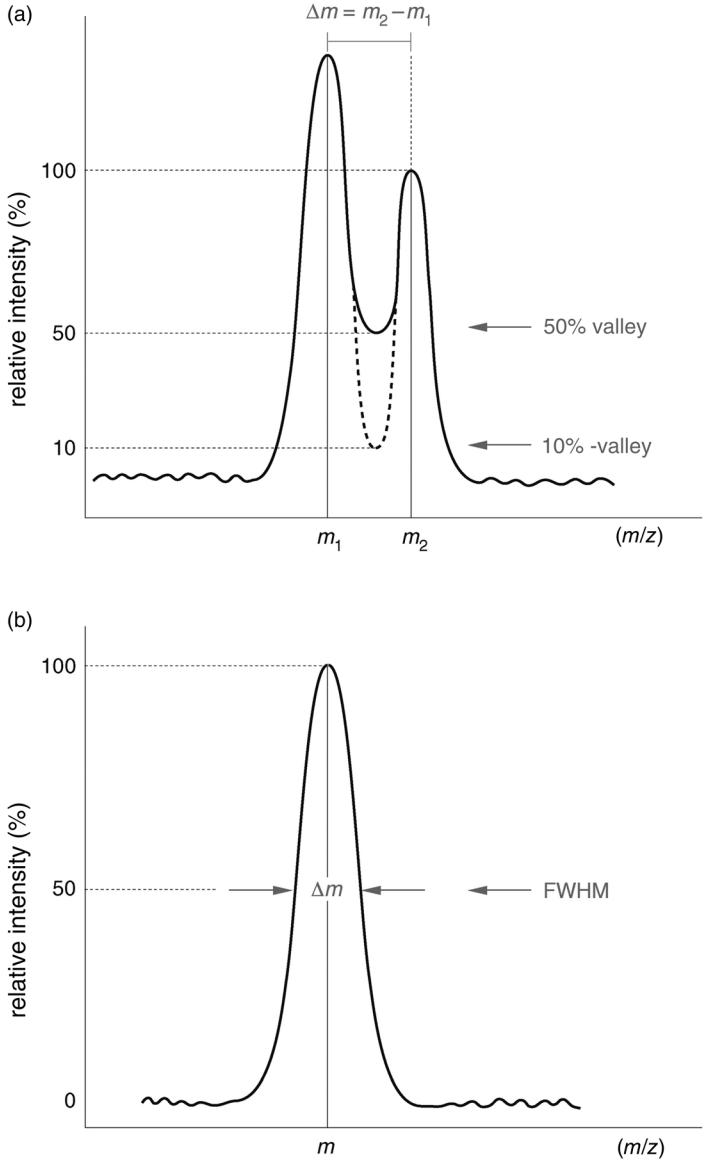


Figure 15.13 Definition of resolution capacity of a mass spectrometer: (a) valley 10% and 50%; (b) full-width at half-maximum (FWHM) of the peak.

An important criterion for mass analyzers is their resolution capacity, which is the ability to separate ions with small mass differences. The resolution capacity R is defined as the ratio of a mass m_1 and the difference Δm in mass between two ions, m_1 and m_2:

$$R = \frac{m}{\Delta m} = \frac{m_1}{m_2 - m_1} \tag{15.3}$$

The separation capacity is different for different mass analyzers. For some analyzers with a relatively high separation capacity certain rules have to be fulfilled; peaks are defined as separated if the peak valley between two peaks represents 10% of the less intensive peak. For quadrupole mass analyzers peak separation is defined by a 50%-peaks-valley definition (Figure 15.13a). A resolution of $R = 1000$ means in this case that ions with a mass of 1001 and 1000 can be separated and detected with a 50%-peak-valley.

Today, the separation capacity is defined using one peak. In this case Δm is defined using the peak's full-width at half maximum (FWHM), which refers to 50% of the maximum peak intensity (Figure 15.13b). In the context of a single unit mass resolution this definition is problematic. Regardless of the resolution of the instrument, the separation of two molecules of, for example, *m/z* 1000 and 1001, is displayed separately in the spectrum. Because both

peaks overlap with their FWHM the signals cannot be separated. Based on the definition of Δm using the FWHM of the peak, separation at baseline is only possible at a resolution of 2000.

Using modern, commercial time-of-flight analyzers a FWHM resolution up to 60 000 is possible. Currently, orbital ion traps have a resolution capacity of 100 000. It is becoming obvious, however, that in terms of resolution future developments will lead to higher capacities. The record holders in terms of resolution capacities are magnetic ion traps, which show resolutions far over 100 000, and FT-ICR MS ($R > 10^6$).

Another important parameter for describing the quality of a mass spectrometer or mass analyzer is the mass accuracy. This parameter gives the accuracy of mass detection, which defines the range in which the measured mass differs from the exact mass. The mass accuracy in this context does not describe the mass precision, which is a parameter of reproducibility. Mass accuracy and mass precision can differ significantly. In most cases the strong variation between these two parameters is due to problems with system maintenance. Typical error sources can be the wrong annotation of signals, overlap of signals, or simply the usage of wrong calibration masses. Both mass accuracy and mass precision are displayed with absolute or relative numbers in parts per million (ppm). A peptide with a molecular weight of 400 Da would be displayed with a relative deviation of 100, 10, or 1 ppm and an absolute deviation of ±0.04, ±0.004, or ±0.0004 Da.

To estimate the performance of a mass spectrometer the comparison of mass accuracy of single measurements is not a good means of quality control. For robust quality control it is reasonable to estimate the median and standard deviation of various independent measurements. When using high resolution mass analyzers mass detection of ±2 ppm should be achieved. Software applications for recalibration of spectra using known masses can lead to a mass precision of ±0.5 ppm. A robust quality control means that out of 1000 measurements a relative error of less than ±6 ppm or ±1.5 ppm can be expected in 997 cases. Only in three cases will the error be higher (99.7% confidence interval).

15.2.1 Time-of-Flight Analyzers (TOF)

One way to measure ions in a high vacuum is by the electronic measurement of time, beginning from the ions in the source up to detection using a time-of-flight (TOF) analyzer. TOF analyzers can be linked directly to MALDI sources. Ions are accelerated using an electrostatic field and reach a kinetic energy of several kilo-electronvolts. After leaving the source, ions pass through a field-free time-of-flight tube and are sorted according to their mass to charge ratio (Figure 15.14a). Separation is possible due to the different velocity of ions that have the same kinetic energy but different mass to charge ratios. Because ions with different *m/z* values, but with the same kinetic energy, reach different velocities, separation depending on *m/z* values is possible. Knowing the acceleration voltage and the flight length of ions in a field-free TOF tube enables determination of the *m/z* values of the ions.

After passing the acceleration voltage U the kinetic energy E_{kin} of ions is defined as:

$$E_{kin} = \frac{1}{2} m v^2 = zeU \tag{15.4}$$

where: m = mass of an ion,

v = speed of the ion after the acceleration distance,

z = charge state,

e = elementary charge,

The speed v results from the time-of-flight t that an ion needs to cross the field-free time-of-flight tube length L:

$$v = \frac{L}{t} \tag{15.5}$$

Insertion of (15.5) in (15.4) results in:

$$\frac{1}{2} m \left(\frac{L}{t}\right)^2 = zeU \tag{15.6}$$

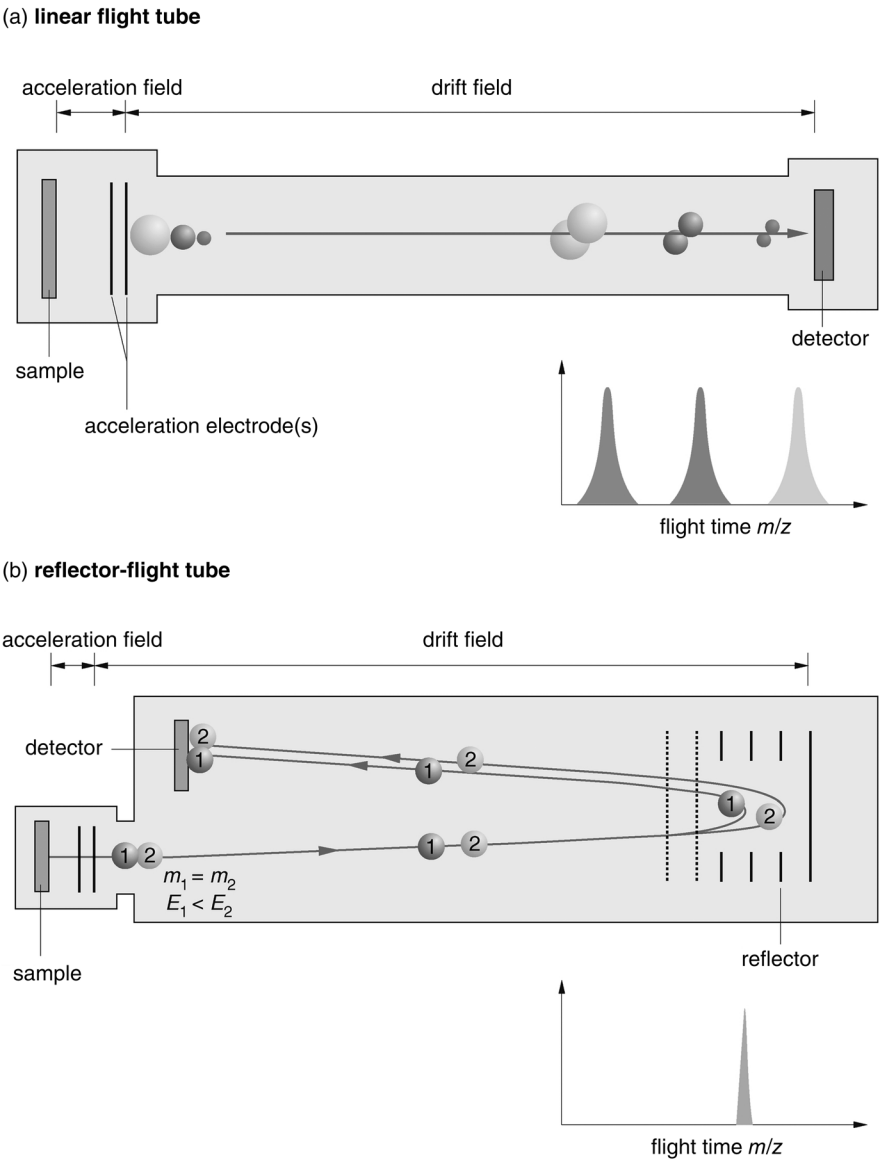


Figure 15.14 Principle of a linear time-of-flight mass spectrometer (a) and of a reflector time-of-flight mass spectrometer (b). The ions generated by the laser impulse have the same charge but different masses (b). The ions generated by laser pulses show the same charge but have different masses and different speed. Heavy ions with high *m/z* values reach the detector later than ions with lower *m/z* values. Ions with the same mass start with different energy distributions, which influences the peak width of the ion signals. By using reflectors the difference in energy distribution can be compensated.

Recalculation to *m/z* results in:

$$\frac{m}{z} = \frac{2eUt^2}{L^2} \tag{15.7}$$

These equations show that the ratio of molecular mass and charge is proportional to the square of the time-of-flight using a TOF mass spectrometer. Therefore, each mass can be determined based on its time-of-flight. Calibration of TOF instruments is based on reference substances with known masses. Typical flight times of MALDI-TOF instruments are microseconds or several hundred microseconds. The size of time-of-flight tubes can be 1 to 4 m.

MALDI-MS is limited in terms of the distribution of the energy of ions during the ionization process. Not all ions are desorbed and ionized at the same time, which results in inaccuracies regarding energy, place, and time. In addition, repelling electrical forces cause a certain distribution of energies during the MALDI process. To prevent premature/early acceleration of the ions different shielding effects can be used. Owing to the delay of ionization in different places of the gas phase some ions do not pass through the whole of the acceleration route. Collisions of the molecules during the acceleration phase cause additional changes in time and energy.

Start errors such as those described above result in different kinetic energy levels of the ions after acceleration in the time-of-flight-tube even though they have the same *m/z* value. Due to the different energy levels ions with the same mass will be detected with small time differences (Figure 15.14a). This leads to broader ion signaling and limitation of mass resolution. When using a TOF analyzer with a linear time-of-flight tube it is possible to reduce the relative energy distribution and its impact on mass resolution by using higher acceleration voltages. But one has to keep in mind that this will also reduce the flight time and, therefore, a more accurate electrical time measurement is needed.

TOF analyzers with ion reflectors provide an elegant solution for this problem (Figure 15.14b). The reflectors change the direction of the ions within an opposing electromagnetic field, which is coupled to the time-of-flight tube. Ions with the same mass but with higher starting energy enter more deeply into the opposing field and pass, consequentially, a longer distance inside the reflector. These ions catch up with slower ions after redirection at a certain point inside the time-of-flight tube. When positioning the detector at this focusing point, ions will be detected at the same time regardless of their different starting energy. The result is a clear signal.

Another option to overcome the limitations of the mass resolution of MALDI is delayed ion extraction. Here, activation of the electrical field above the sample surface is delayed in relation to the desorption laser pulse. Ions with a higher starting energy have to travel a greater distance to the sample surface during the delay, time which leads to lower kinetic energies of those ions after activation of the electric field compared to ions with a lower starting energy. Choosing appropriate values for the delay time and extraction field strength enables compensation of the influence of different starting energy distributions on the time-of-flight behavior of ions at the detector. In practice the improvements for resolution, delayed ion extraction, and usage of reflectors complement each other. The combination of these techniques improves the mass resolution up to 60 000 (FWHM).

The main advantage of TOF mass spectrometers is that the induction and identification of low ion current is possible with a single laser shot due to the TOF's high ion transmission rate. In theory the mass range of a TOF mass spectrometer is not limited but in practice there are some limitations. The detectors are limited regarding resolution and sensitivity for high masses.

Initially, it was difficult to combine TOF analyzers and ion sources like ESI. The results were intensity loss and poor peak resolution of spectra. These problems were overcome by using orthogonal ion extraction (ESI-o-TOF-MS) (Figure 15.12). Using ESI-o-TOF-MS, continuously emitted ions of ESI sources were transferred into ion packages using orthogonally orientated electrical fields and transmitted into the TOF. Since this time the special advantages of TOF can be used for ESI-MS. For MALDI-MS the usage of orthogonal transmitted ions is possible, too (MALDI-o-TOF-MS). Figure 15.24 below shows the schematic composition of a TOF mass spectrometer as a part of a hybrid instrument.

Today both options are used, but the TOF analyzer with direct ion extraction is the most prominent MALDI mass spectrometer. This is based on the sensitivity of direct ion extraction and less transmission losses, which lead to shorter analysis times. In some cases direct ion extraction is described as linear extraction to differentiate direct ion extraction from orthogonal ion extraction. Using this terminology one has to keep in mind that TOF analyzers are also differentiated according their geometries as linear and reflector instruments.

The main advantage of orthogonal ion extraction is that a high vacuum is not needed; the instrument can work under atmospheric pressure conditions. Therefore, it is possible to remove excess energy using internal cooling by collision with other molecules. Especially, unstable ions (e.g., RNA or DNA molecules) can be stabilized effectively. If the internal energy of the MALDI instrument is too high, molecules fragment very rapidly so that time-of-flight analysis is no longer possible.

15.2.2 Quadrupole Analyzer

The quadrupole mass spectrometer is basically a mass filter. This means that under a given physical condition only ions with a specific *m/z*-ratio will be transmitted to the detector. This is achieved by an array of four parallel rod-shaped metal electrodes (quadrupole). Ions of a defined *m/z*-ratio can travel through the quadrupole on a stable oscillating track under the influence of a

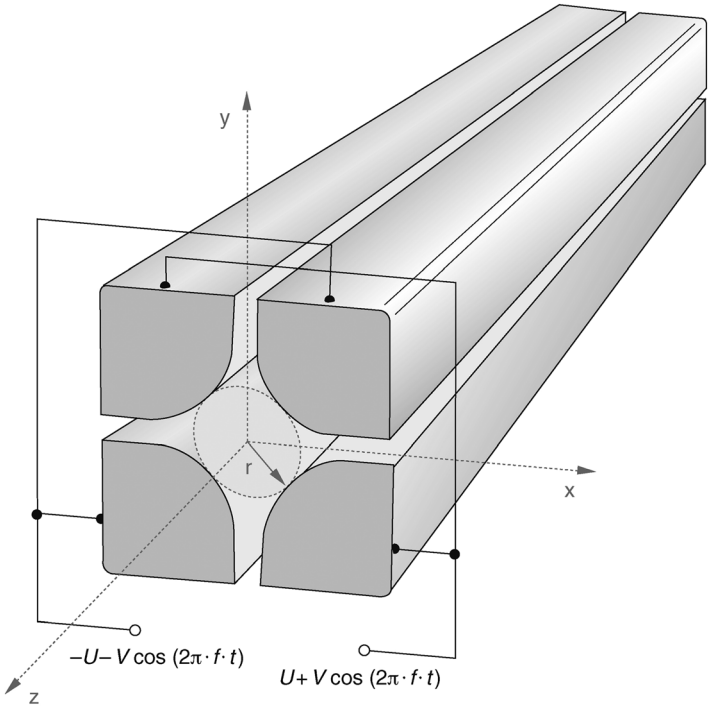


Figure 15.15 Electrode shape and electrode arrangement in a quadrupole mass filter.

combined AC and DC field (Figure 15.15). All other ions with different *m/z*-ratio fly on unstable tracks and are stopped by collisions with the metal rods.

As shown in Figure 15.15 a quadrupole consists of four hyperbolic formed rod-shaped metal electrodes that are arranged on a circle of radius *r* about the *z*-axis (circular cylindrical rods are used for reasons of cost). The rods are set with a DC and AC voltage *U* and a frequency *f* (*V* cos (2π*ft*)). Opposite rods are of the same polarity of DC voltage and the same phase of AC voltage. Adjacent rods have an opposite polarity and a phase offset of 180° while an electric potential ϕ is produced around the *z*-axis:

$$\phi(x, y, t) = [U + V\cos(2\pi f t)]\frac{x^2 - y^2}{r^2} \tag{15.8}$$

The ions obtained by a low acceleration voltage of 10–20 V have sufficient translational energy and move in the direction of the *z*-axis into the electric field of the quadrupole.

The movement of the ions in the *xy*-plane is described by the Mathieu equations:

$$\frac{d^2x}{d(\pi f t)^2} + [a + 2q\cos(2\pi f t)]x \tag{15.9}$$

$$\frac{d^2y}{d(\pi f t)^2} + [a + 2q\cos(2\pi f t)]y \tag{15.10}$$

The relationship between a transferred ion of mass *m* and the elementary charge *z* is described by the parameters *a* and *q:*

$$a = \frac{2zeU}{m(\pi f r)^2} \quad \text{and} \quad q = \frac{zeV}{m(\pi f r)^2} \tag{15.11}$$

The radius *r* between the rods is important, as is the electric field consisting of the DC voltage *U* and the AC voltage *V* with the frequency *f*.

The Mathieu equations have two types of solutions: One solution results in finite amplitudes of the oscillations corresponding to a stable movement through the quadrupole. The other solution leads to amplitudes that grow exponentially in the *x*- and/or *y*-direction. For ions with a given *m/z* ratio there are certain values for the parameters *a* and *q*. Here stable oscillations in the *x*- and *y*-direction are possible. Numerical evaluation of the Mathieu equations yields a stability diagram for *a* and *q* (Figure 15.16). Each point in this diagram represents ions with a specific

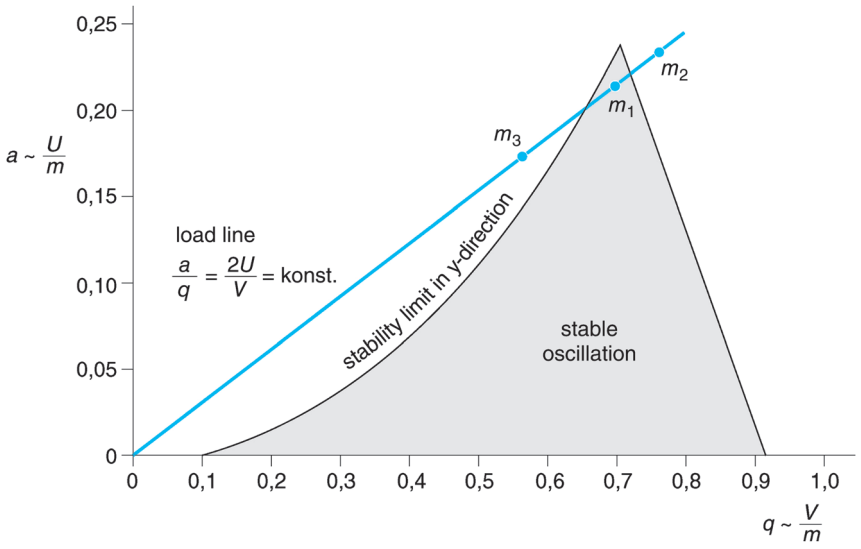


Figure 15.16 Mathieu equations stability diagram for the two-dimensional quadrupole field in the *x*- and *y*-direction.

m/*z*-value with given values for *r*, *U*, *V,* and *f*. From the equations for the parameters *a* and *q* it is found that the ratio *a*/*q* is always equal to 2*U*/*V*:

$$\frac{a}{q} = \frac{2zeU}{m(\pi fr)^2} \frac{m(\pi fr)^2}{zeV} = \frac{2U}{V} \qquad (15.12)$$

This will be the same for all masses on the straight line *a*/*q* = constant, the so-called working load line.

The voltage *U* and the amplitude *V* of the alternating field are increased at the same time when scanning of the mass range occurs. In this situation the ratio *U*/*V* – and thus the ratio *a*/*q* – and the frequency *f* (radio frequency range) are kept constant. Ions of different masses are successively brought into the stable region of the quadrupole field. In principle, an increase in the frequency while keeping *U* and *V* constant would be possible, but this is difficult to realize due to technical reasons. The mass *m* is directly proportional to *U* and *V,* so that with an increase of voltage a linear mass spectrum is obtained. The mass separation and thus the mass resolution $m/\Delta m$ are controlled by the ratio of *U* to *V*. With this chosen ratio the ions with a particular *m*/*z* ratio reach into the range of stable oscillation (see intersections of the working load line in the stability diagram, Figure 15.16). The ions with m_1/z in this example are due to a stable oscillation in the *x*- and *y*-direction on a trajectory through the quadrupole to the detector. Under these conditions all other ions in the *x*- or *y*-direction (e.g., m_2/z, m_3/z) undergo unstable oscillations, so that they are stopped by collisions at the rods or lost between the rods.

In the ideal case of a technically perfect quadrupole the mass resolution $m/\Delta m$ depends only on the ratio *U*/*V*. The maximum mass resolution would be achieved theoretically when the working load line in the stability diagram touches the tip of the region of stable oscillations. This would correspond to a value of 0.1678 for the ratio *U*/*V*. However, in practice the achievable mass resolution depends on the initial velocity of the ions in the *x*- and *y*-direction and on the deviation of ions from the ideal *z*-direction when entering the quadrupole. The sensitivity decreases with increasing resolution because ions with a specific *m*/*z* ratio do not reach the detector.

Commercial quadrupole analyzers have a maximum mass range of up to *m*/*z* = 4000 and achieve resolution values between 500 and about 5000. Here, the mass resolution can be varied while an increased resolution is associated with a loss of sensitivity. In general, the devices are set to a nominal mass resolution that results in a complete accessible mass range. The ions are usually detected by a secondary electron multiplier (Section 15.3.1). Quadrupole mass spectrometers are characterized by a high ion transmission from the source to the detector. They are easy to focus and calibrate and have excellent calibration stability during continuous

operations. These practical advantages have led to the widespread use of quadrupole machines in organic and biochemical analyses. To record a spectrum (scanning the quadrupoles across the mass range) two parameters are important: first, the step size into which the selected scan area is divided (in practice between 0.05 and 0.5 amu) and, second, the time (dwell time) for the step measurement (lower millisecond range).The step size affects the accuracy and the dwell time affects the sensitivity of the mass determination. Thus, it is preferable to work with the smallest possible step size and the longest possible dwell time. The product of step size and dwell time gives the required scanning time to record a spectrum (usually several seconds). The analyte is continuously supplied during the measuring time for a spectrum. Therefore, the scan time determines the absolute consumption of the analyte. In practice a compromise is made, according to the predetermined amount and concentration of the analyte, between step size and dwell time. For example, a spectrum could be obtained of a peptide in a measurement range of 300–2000 amu with a step size of 0.2 amu and a dwell time of 0.7 ms in 7 s. Here, particularly sensitive measurements can be performed when the mass of the analyte is known. It is scanned only over the limited range of the expected molecular ion (about 5 amu) and the dwell time is set in the upper millisecond to second range. For the above selected peptide a dwell time of 0.5 s for a measuring range of 5 amu would be reached by using the same step size (0.2 amu) and the same scan time (7 s). This so-called single ion monitoring (SIM) scan mode is not limited to quadrupoles and is often used for the highly sensitive detection of known substances in mixtures.

15.2.3 Electric Ion Traps

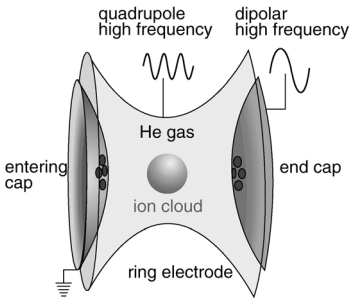


Figure 15.17 Schematic design and operating principle of a three-dimensional electric ion trap.

As an alternative to the quadrupole analyzer a three-dimensional electrical ion trap can be used for mass analysis for MALDI and ESI ions. This principle is based on the trapping of ions in a suitable electric field. The ions may be held for variable times (microseconds to seconds) in stable orbits and then analyzed according to their masses. A three-dimensional electric ion trap consists of a ring electrode and two end caps that are set with AC voltages (Figure 15.17). In the middle of the end-caps there are small, centric openings for the inlet and the discharge of ions. The entire analyzer is not greater than a cube with ca. 10 cm edge lengths. The function of the ion trap and of the quadrupole are largely based on the same principle. In both mass analyzers the solution of the Mathieu differential equations determines the value ranges of applied DC and AC voltage, in which ions describe stable orbits. The ring electrode of the ion trap can be seen as a curved quadrupole rod whose ends are connected to each other. The end caps define the system from both sides, like two opposing rods in the quadrupole. The ions are now no longer transported in the quadrupole over a certain distance and describe closed tracks in the system, that is, they are "trapped". A stability diagram is formed by the overlap of the stable regions of the two end caps and the ring electrode (Figure 15.18). The aim here is to trap ions over a wide mass range and to store them. Considering the stability diagram, the widest mass range for stable ions is just reached if $a = 0$ and, thus, no DC voltage is applied. Therefore, ion traps are usually operated only with AC voltage. By applying the AC voltage to the ring electrode a quadrupole field is formed inside the trap. This field generates a spatially extended potential gap, which fixes the ions in the inner part. The ions entering the trap are captured by the electric field only to a very small extent. The ions enter with a certain speed because they were produced in an external source outside the trap. Similarly to a marble that is rolled from a distance into a small gap and rolls out again due to excessive speed, most of the ions would "run up" the potential gap due to their entry speed and leave the trap or crash against the surrounding walls. Therefore, to stop the ions after they enter the trap, the trap is filled with helium at a pressure of 3×10^{-6} bar. By collisions with the helium atoms the ions reduce their speed and can be captured by the quadrupole field efficiently.

During a measurement cycle the ions are accumulated in the trap only for a limited time, usually between 0.1 and 10 ms. After this process the transfer line between the ion source and trap is blocked by changing the applied electric potential, which means that no further ions can enter the trap. This is to avoid a too high charge density of ions in the trap, which would result in mutual repulsion of the stored molecules and to an incorrect determination of ion masses. Any influence on the mass analysis by subsequent ions should also be avoided. For detection, ions of increasing molecular weight are ejected out of the trap. The ion exit

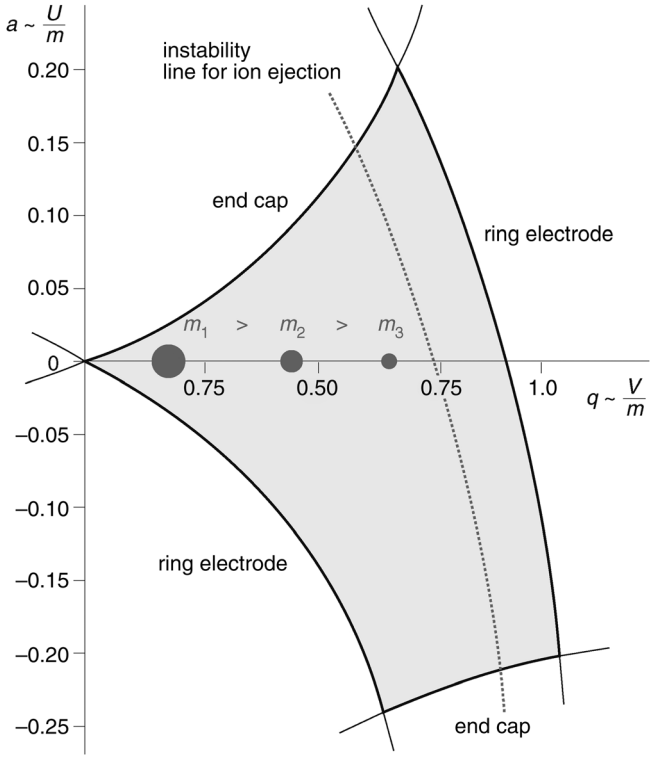


Figure 15.18 Stability diagram of the Mathieu equations for a three-dimensional electric ion trap.

can be carried out in two ways. The simplest way is (similar to the quadrupole) to increase the AC voltage amplitude V (q) to force out the ions successively from the stability region (Figure 15.18). However, this would allow only slow scan speeds, similar to the quadrupole, and the achievable mass resolution would be quite limited. A far better alternative is to eject the ions from the trap by using multipole fields. For this purpose, the coupling of the quadrupole field at the ring electrode is used with a dipolar field that is generated at the same time at the end of the caps. The coupling of these two fields is caused by a specific geometry of the ring surface or the end caps of the ion trap. By additive superposition of the two fields a plurality of fields of a higher order, so-called hexa-, octa-, and decapole-fields, are generated in the trap. With increasing q the ions can absorb energy at many different multipole fields. A resonant excitation turns them extremely quickly into large oscillations, so that they are ejected within a very short time from the trap. In the context of Figure 15.18 it is equal to the sharp instability lines that lie within the actual stability region. The ions are ejected from the trap once they reach this sharp line of instability with increasing q. In commercial ion trap analyzers the latest generation of scan speeds can achieve up to $26\,000\,\mathrm{u\,s^{-1}}$ by applying the multipole effect. This is more than 20 times faster than in quadrupole systems. Here, a higher repetition rate for the recording of spectra is important for a high so-called duty-cycle (the rate of the actually detected ions from a continuously emitting ion source during a measurement cycle).

At the same time the resolution was significantly improved to values above 10 000. The maximum m/z-range can be up to 6000. In addition to the *three-dimensional,* there are also *two-dimensional electric* traps, so-called linear ion traps. In these traps the ions are stored in a two-dimensional electric quadrupole field. Owing to the linear geometry of the trap, space charge effects are reduced. As a result, a larger number of charged particles can be injected and stored in the trap. Thus, the detection sensitivity can be increased.

15.2.4 Magnetic Ion Trap

Magnetic ion traps contain a strong homogeneous magnetic field that forces the ions on a circular orbit perpendicular to the magnetic field. This movement is called cyclotron-movement

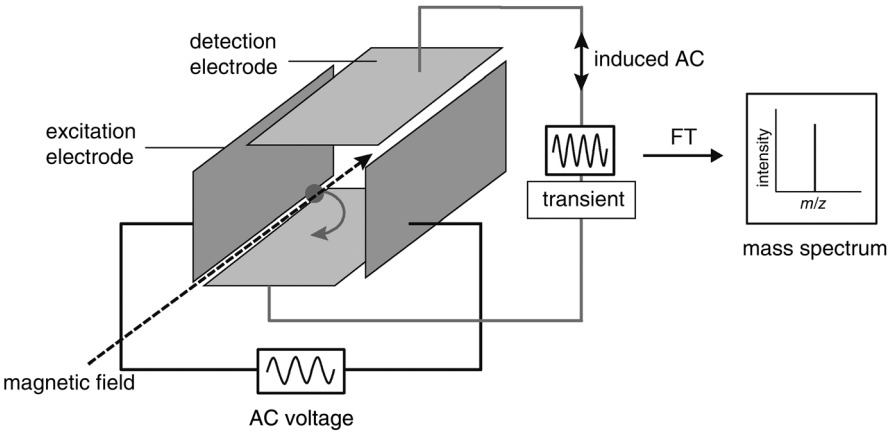


Figure 15.19 Schematic design and operating principle of a magnetic ion trap.

or cyclotron-oscillation. Their frequency (cyclotron-frequency) is inversely proportional to *m/z* and directly proportional to the strength of the magnetic field:

$$\omega_{ion} = \frac{zeB}{2\pi m} \tag{15.13}$$

where: z = the number of charges,
e = the charge of an electron,
B = the magnetic field strength,
m = mass.

This movement can be detected by an image current (mirror current), which is induced by the circulating particles on an electrode/plate detector. For circulating ions with different *m/z*, their cyclotron frequencies can be determined by a Fourier transformation of the induced image current signal. The result is a cyclotron frequency spectrum that with Equation 15.13 can be converted into a mass spectrum immediately. This is the underlying principle of a magnetic ion trap, better known as a Fourier transform ion cyclotron resonance mass spectrometer (FT-ICR-MS) (Figure 15.19). The movement of the ions in the trap is limited by the magnetic field in a perpendicular direction and by an electrostatic trapping potential. At the beginning of the measurement the trapped ions circulate close to the center of the trap. The ions are excited by external excitation electrodes with a pulsed alternating electric field; the electrodes are parallel to the magnetic field. When the cyclotron frequency and the frequency of the incoming AC electric field match, the resonance case occurs (cyclotron resonance) and the cyclotron radius of the ions increases by absorbing energy from the alternating field. This process is precisely timed. This means that the corresponding alternating field is present until the ion cyclotron radius approaches half of the distance to the detector plates. The ions then fly very close to the plates and induce a detectable image current. By overlying different frequencies and their rapid temporal variation it is possible to activate all ions, or only a specific selection of them, and analyze them. The resolution of an FT-ICR-MS increases with the strength of the magnetic field used. This is generated by superconductor magnets and reaches with present-day devices field strengths of up to 15 Tesla, which corresponds to 300 000 times the strength of the earth's magnetic field. The resolution is very high and can reach values in the range of several 100 000 to 2 million. This allows mass accuracies in the ppb (parts per billion), but requires analysis times of several seconds to minutes.

15.2.5 Orbital Ion Trap

The third option of an ion trap is to move the ions around a central spindle-shaped electrode on a stable orbit (orbitals). This principle is used in the orbital ion trap, usually known as the "Orbitrap" (Figure 15.20). The applied electrostatic field and the resulting force of attraction towards the central electrode compensate exactly the centrifugal force, so that the ions move like satellites in an orbit. Electrodes on the end caps produce a potential barrier that prevents the ions leaving the Orbitrap laterally. In addition to the circular motion around the central

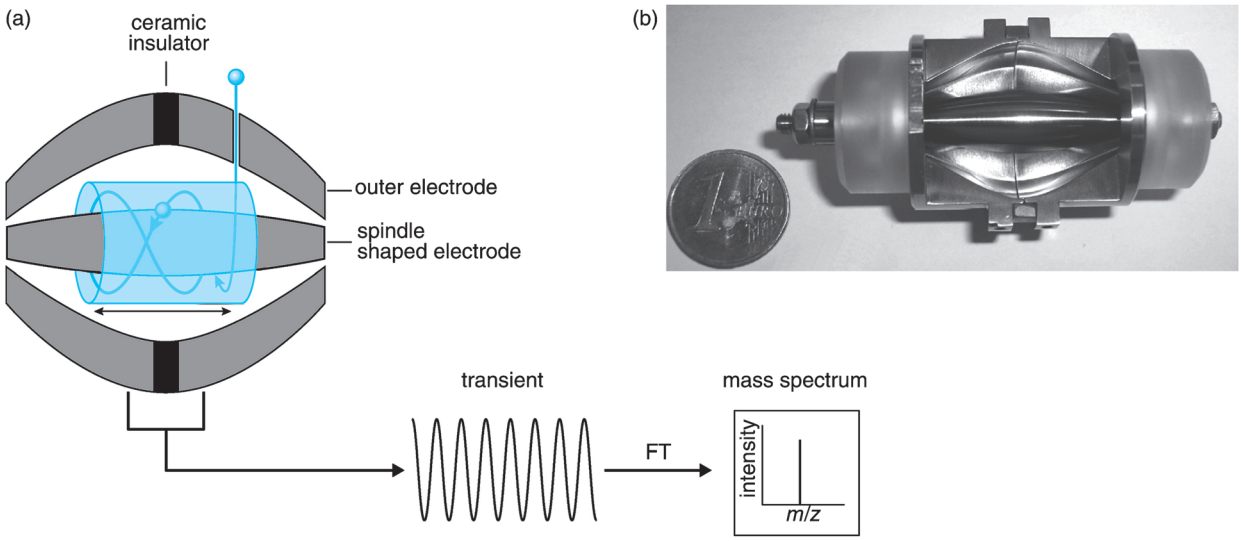


Figure 15.20 (a) Schematic design and operating principle of an orbital ion trap. (b) The Orbitrap. Source: with permission of Thermo Fisher Scientific.

electrode, lateral (radial) oscillations occur. They are independent of entry angle and velocity of the ions and depend only on their *m/z* ratio. Here, an outer electrode that is divided in the middle by a ceramic ring can measure these oscillations as an induced current. In the simplest case, when only one type of ion is present in the trap with a fixed *m/z* ratio, a sinusoidal oscillation is generated. From the frequency of this oscillation the *m/z*-ratio can be calculated:

$$\omega = \sqrt{\frac{k}{m/z}} \tag{15.14}$$

where ω is the frequency of oscillation and k is the instrument constant. When different ions are present with different *m/z*-ratio, a superposition of different sinus oscillations, the so-called transient, is obtained.

Here the Fourier transform is used to calculate a mass spectrum. Although the basic principle of a trap with ions in orbits had long been known it was realized in practice only a few years ago. The problem was mainly the injection of ions into the trap. For a continuous ion current it is not possible to trap the ions in stable circular orbits, because they would fly by the central electrode due to their high speed. To solve this problem ions are collected in an upstream C-shaped ion trap (C-trap) and injected into the Orbitrap as bundled packages. During the presence of the ions the voltage at the central electrode is briefly lowered to decelerate the ions. The Orbitrap enables measurements with a mass accuracy of 0.5–2 ppm and a mass resolution of up to 100 000 FWHM. The Orbitrap does not achieve the resolution of magnetic traps, but it runs without an expensive, high-maintenance superconducting magnet. In addition, the analysis times are shorter for the same resolution. This means that the Orbitrap analyzes the trapped ions faster than a magnetic ion trap and can analyze more samples per time unit.

15.2.6 Hybrid Instruments

Especially in biological mass spectrometry hybrid instruments or tandem mass spectrometers that combine several mass analyzers are widely used. For example, one analyzer can be used for fragmentation, the other for analysis. In the following the different characteristics of mass analyzers are highlighted and different hybrid instruments are described.

Triple-Quadrupole (Triple-Quad) Triple quadrupole instruments were the first instruments to enable not only the separation of ions regarding their mass to charge ratio but also the isolation of single ions for fragmentation. Fragmentation of ions is necessary to gather information about their structure. Triple quadrupole mass spectrometers consist actually of four quadrupoles (Q0 to Q3) with two of them for scanning (Figure 15.21). Q0 is a supportive

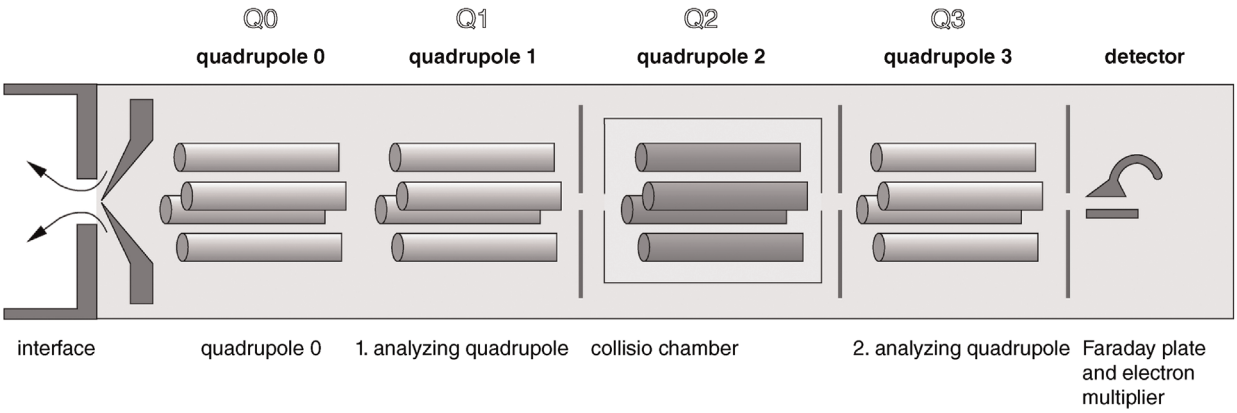


Figure 15.21 Schematic assembly and functional principle of a triple quadrupole.

quadrupole driven by alternate current without any direct current components. Using Q0, ions are focused and transferred to the central ion path. Owing to the absence of stray fields caused by direct current components, ions are efficiently transported into the quadrupoles' center and start oscillating. The first analyzing quadrupole is Q1 where ions are scanned. Q2 transfers ions through a collision chamber flushed with an inert gas (e.g., nitrogen, helium, argon). Collision of the ions with the inert gas molecules or atoms induces fragmentation of the ions in Q2. In Q3 the corresponding molecular masses of the fragment ions are analyzed. The produced fragment spectra are referred to as tandem mass spectra or MS/MS or MS^2 spectra. Using a triple quadrupole mass spectrometer different types of MS/MS analyses are possible (Figure 15.22).

Identification, Section 16.6.1

Product Ion Analysis For product ion analysis Q1 is configured to transfer only ions with defined *m/z* values. These precursor ions are fragmented in Q2 and product ions are analyzed in Q3. The analysis of product ions is the most common MS/MS method because identification or quantification of single components of complex mixtures is possible.

Quantification, Section 15.8

Precursor Ion Analyzes The reversed allocation of precursor and product ions is possible as well. Therefore, ions of the whole mass range are transferred to Q1. Then the ions reach Q2 sequentially and are fragmented.

If Q3 is pre-set for a certain *m/z* value Q1 transfers the corresponding precursor ions of the fragment ions into the collision chamber. The *m/z* range of the precursor spectrum displays the *m/z* range of Q1. Using the precursor ion analysis method modifications like glycosylated peptide ions of molecule ions can be detected by searching for specific fragments for the modifications such as carbohydrate fragments.

Neutral Loss Scan The neutral loss scan is a method for detecting the loss of neutral elements during fragmentation. Similar to precursor ion analysis, Q1 analyses all masses and Q2 is used for fragmentation. Q3 is synchronized with Q1 during the scanning mode and scans all masses as well but in this case with a mass offset that correlates with the mass of the specified neutral loss. Only ions with a correlating precursor ion in Q2 that produce the selected mass difference are analyzed at the detector. The neutral loss scan is a useful technique for highly sensitive identification of molecules with known masses in a complex mixture.

SRM- and MRM-Analysis To analyze biological samples using mass spectrometry triple quadrupole mass spectrometers are commonly used for single reaction monitoring (SRM) or selected reaction monitoring or multiple reaction monitoring (MRM). For many years SRM/MRM has been used routinely in chemical trace analysis. This technique enables highly sensitive detection of known molecules in a highly complex mixture containing molecules that are very similar to each other.

Verification, Section 15.6.2

The molecule with known mass to be analyzed is selected in Q1 and fragmented in Q2. In Q3 only a selective mass range is analyzed and only one or several fragment ions with known *m/z* values (constant *m/z*) is scanned. The maximum detection sensitivity with SRM is achieved by using only one fragment ion. The error rate is minimized when several fragment ions are analyzed.

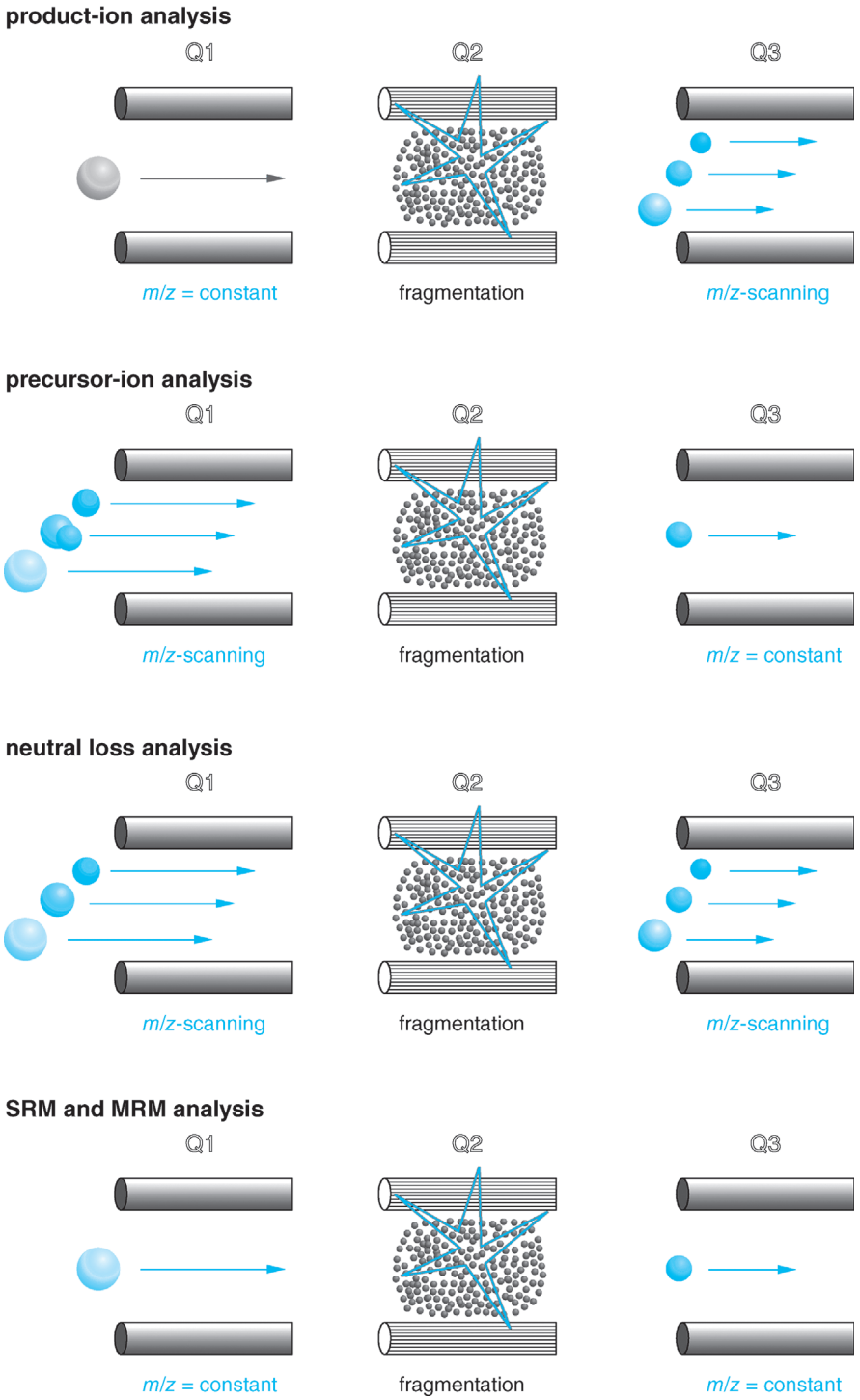


Figure 15.22 Different principles of triple-quadrupole mass spectrometer analyses.

For MRM alternating time frames are used in Q1 to access very small *m/z* ranges, and for the detection of a molecule with a specific *m/z* value known fragment ions for identification can pass to Q3. MRM is a compromise between detection sensitivity and the number of analytes.

Tandem-TOF (TOF-TOF) Tandem TOF-TOF instruments can be used for controlled fragmentation of MALDI ions. Figure 15.23 shows a TOF-TOF instrument with its collision cell in between the two independent analyzers. After ion formation in the MALDI source ions are transferred and accelerated into the first linear time-of-flight tube (TOF1).

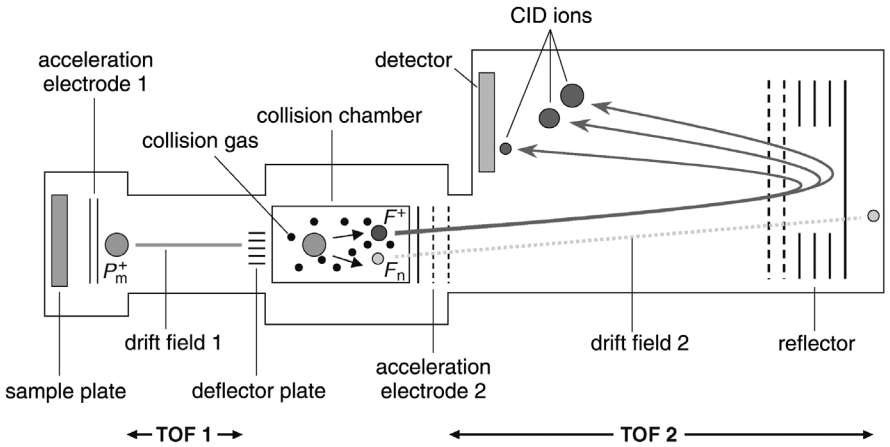


Figure 15.23 Schematic assembly and functional principle of a MALDI TOF-TOF analyzer. In TOF1 molecule ions (P_m^+) are selected based on their *m/z*-value and are fragmented using argon gas molecules in a collision chamber. The resulting fragment ions are analyzed in TOF2.

Selection of the ions based on *m/z* values takes place in TOF1. The selected masses are decelerated and are fragmented inside a collusion chamber filled with inert argon gas. The fragment ions leave the collision chamber, are accelerated, and analyzed in TOF2. The great speed and the possibility to run MS/MS analysis at the same time are important characteristics of TOF-TOF instruments. More than a thousand MS/MS spectra can be produced per second and are often averaged to improve detection sensitivity. Besides the described TOF-TOF analyzers, another alternative is possible that lacks the collision chamber between the TOF analyzers. This analyzer was developed and optimized for the analysis of fragment ions that are formed in TOF due to the surplus energy they absorbed while passing the source. The ion fragments move with the same speed as the ions before decay, therefore local separation in a linear TOF is not possible. Such decay events are typical for MALDI.

After separation in TOF1 selected ion molecules and corresponding fragment ions are isolated and accelerated before entering TOF2. The ions and fragment ions now have a higher electrical potential and are further accelerated. The different *m/z* values of ions and fragment ions result in different speed maxima and therefore separation in TOF2. In principle this function is similar to an elevator and is described as LIFT.

Prompt and Metastable Decay (ISD,PSD), Section 15.4.2

Quadrupole-TOF (Q-TOF) The Q-TOF is a hybrid mass spectrometer combining the power of quadrupole analyzers (easy selection of precursor ions and effective fragmentation) and a reflector time-of-flight analyzer (unlimited mass range, high resolution capacity, high identification efficiency, and very short analysis times). Like a triple-quadrupole instrument, Q1 in a Q-TOF-MS works as a mass filter for precursor ion selection and Q2 is the collision cell for the generation of fragment ions. Mass analysis works differently, instead of using Q3 the mass analysis is performed in an orthogonal-reflector-TOF-MS (Figure 15.24).

Besides the analysis of product ions it is possible to perform precursor ion scans using a Q-TOF-MS. For precursor ion analysis, ions are selected in Q1 and transferred into Q2, starting with the lowest *m/z* values. For all selected precursor ions the corresponding product ions that are generated in Q2 are analyzed with TOF and MS/MS spectra are produced.

With suitable software it is possible to assign product and precursor ions of interest; the corresponding precursor ions scans are reconstructed from the scanned product ion spectra. It should considered out that the precursor ion analysis using a Q-TOF-analyzer is slower and less sensitive compared to a triple quadrupole MS instrument.

Q-TOF instruments are commonly combined with nano-ESI but a combination with MALDI is possible as well. Since MALDI produces mainly single charged peptide ions inert argon gas is used as collision gas. Owing to the relatively high collision cross-section of argon the inner energy of MALDI ions is effectively increased, inducing extensive fragmentation.

Linear Ion Trap with Magnetic Trap or Orbitrap In mass spectrometry magnetic and orbital ion traps are commonly coupled to linear electric ion traps (Figure 15.25). The particular advantage of this set-up is that the individual strength of the different mass analyzers complements each other. The special strength of linear ion traps is their high capacity, which means that a huge number of ions can be collected for the subsequent analysis. Moreover, linear ion traps are

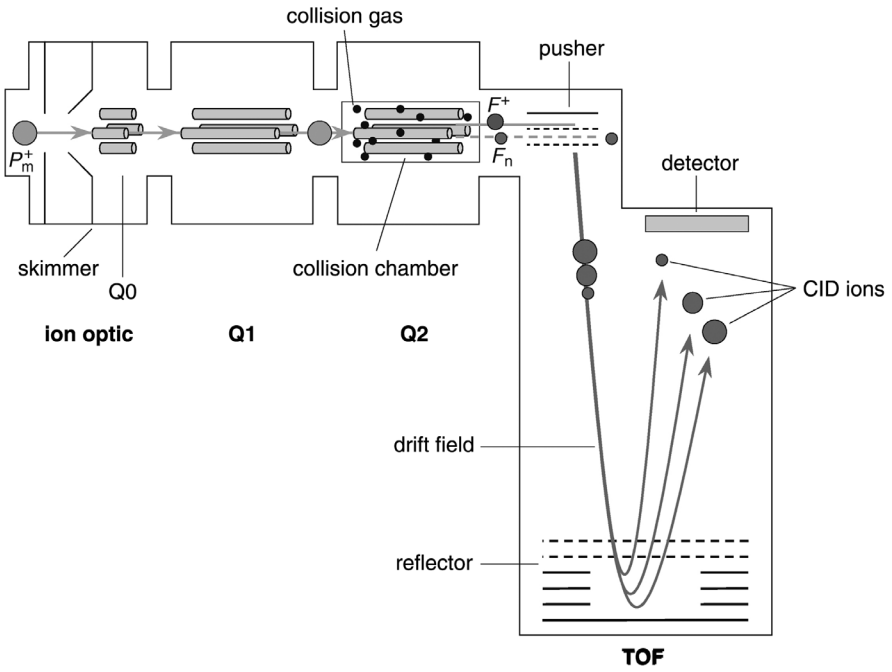


Figure 15.24 Schematic assembly and functional principle of a quadrupole time-of-flight mass spectrometer (Q-TOF-MS). To generate MS/MS spectra using a Q-TOF mass spectrometer, ions (P_m^+) of a selected *m/z* value are selected in Q1 and fragmented in Q2 using an inert collision gas. Mass analysis of fragment ions takes place in the reflector TOF analyzer.

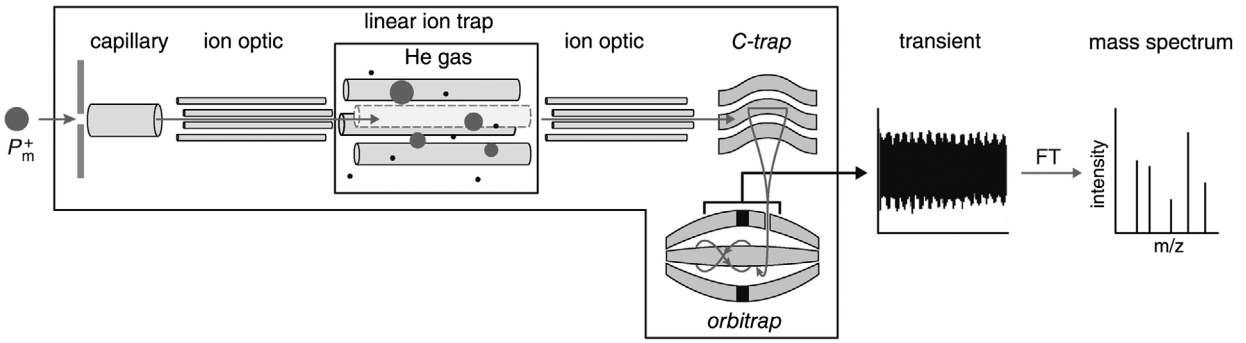


Figure 15.25 Schematic assembly and functional principle of a hybrid mass spectrometer LTQ-orbitrap consisting of a linear electric (LTQ-) and an orbital ion trap.

unbeatable in terms of precision, speed, and efficiency of ion fragmentation. Magnetic traps have a lower capacity and exhibit a slower ion processing time. The commercially available orbital traps today cannot perform MS^n-analyses. The special advantages of both analyzers are their high resolution capacity and the resulting low error rate during mass analysis.

Another advantage of combining two kinds of ion traps is that both analyzers can be run at the same time. A typical example of this experimental set up is the measurement of MS/MS spectra in the linear ion trap when, meanwhile, the magnetic or orbital ion trap collects the following MS spectrum. When choosing a high resolution analyzer for your experimental set up it is important to be aware that magnetic ion traps have a higher resolution capacity while orbital traps have a higher speed during scanning and lower cost of purchase and maintenance.

15.3 Ion Detectors

To detect ions in mass spectrometers special detectors have been developed. They all differ, depending on their requirements, in their principle of operation and in their performance. The most important ones are briefly presented.

15.3.1 Secondary Electron Multiplier (SEV)

The secondary electron multiplier (SEV) is a device that makes it possible to amplify the smallest electron currents or even single electrons with high time resolution. One electron that hits a metal or semiconductor surface can introduce up to ten free secondary electrons, which are then accelerated by electric fields. This procedure is repeated continuously. Owing to the exponential amplification of these events a measurable current can be detected.

To detect large singly or doubly charged molecular ions the SEV can be combined with a conversion dynode that is able to post-accelerate the ions before the impact. The aim is to increase their speed to the extent that secondary electrons are released. The ions with increasing m/z, secondary electrons, and increasingly small secondary ions (m/z <100 Da) are detached from the dynode (ion/ion secondary conversion). The different masses of the secondary ions generated at the conversion dynode lead to a delay dispersion between the conversion dynode and the SEV. This effect causes, beside a narrow secondary electron signal, an additional much wider secondary ion signal. For this reason conversion dynodes are used only in combination with MALDI for the detection of very large molecules, for example, antibodies. In these cases, the ion/ion secondary conversion is in the focus of the detection. The related signal broadening is not of relevance because the isotopic pattern is not resolved.

Constructions with Discrete Dynodes For this purpose several dynodes, specially shaped electrodes, are connected in series. An accelerator voltage is applied to the entire series of dynodes. This occurs in such a way that the electrons are accelerated from dynode to dynode. Since a dynode can emit and absorb electrons it serves perfectly as a cathode and likewise as an anode (Figure 15.26).

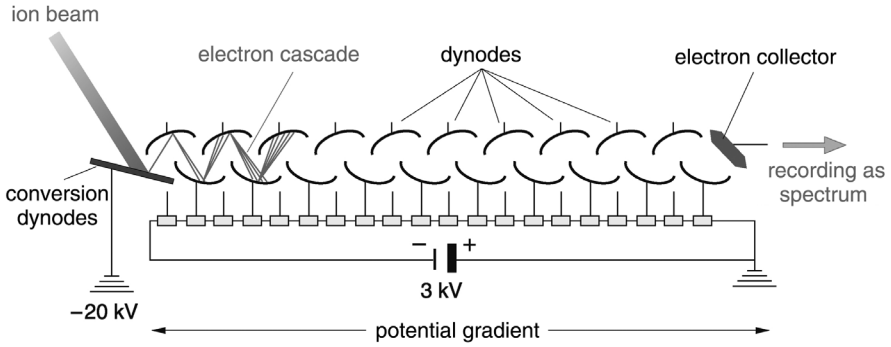


Figure 15.26 Schematic structure and function principle of a secondary electron multiplier (SEM) with series of dynodes and the related conversion dynode.

Channel Electron Multiplier A channel electron multiplier (CEM) generates about 10^8 electrons from a single ion. It often consists of a glass tube whose inner surface is coated with a high electrical resistance layer. The resistance between the cathode at the open end of the tube and the closed end of the anode is approximately $10^9\ \Omega$; the length to diameter ratio is typically around 70. In operation a voltage of about 2 kV is used. The architecture of the detector prohibits deeper penetration of the tube. During the impact released secondary electrons are accelerated towards the opposite wall. On reaching the wall they release secondary electrons. This process is repeated until the electron avalanche reaches the anode.

Microchannel Plates A microchannel plate (MCP) is a surface based image resolution secondary electron multiplier. It is used for low noise amplification of low ionic current flows. Inside the MCP device the usual single channel is substituted by a microchannel plate that contains several thousand very small, paralleelly arranged channels (Figure 15.27).

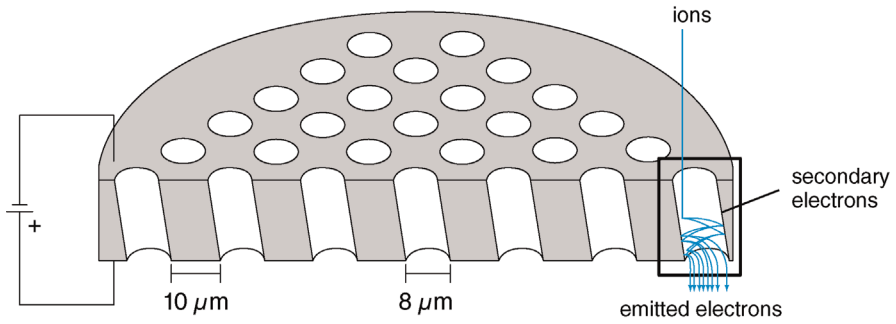


Figure 15.27 Schematic design and operating principle of a microchannel plate.

15.3.2 Faraday Cup

The Faraday cup (also called the Faraday detector) consists of a metal cup. The cup collects all ions and compensates their charge with electrons from a connected high impedance resistor. This process is detected by the change in voltage, which in this case reflects the ion current (Figure 15.28).

In other words the Faraday cup gives information on the number of trapped charges per unit time. The shape of the detector (cup) and also optionally additional electrodes prevent the release of possible secondary ions out of the cup. In this way measurement of the current remains correct. Faraday cups are used as an alternative or in addition to SEV. The advantages of the Faraday cup are reliability, robustness, and most of all the possibility to measure the absolute ion current. In addition, the sensitivity is constant in time and does not depend on the *m/z* values of the ions (such as in SEV). The disadvantages compared to SEV are the significantly lower sensitivity and longer reaction time. They depend on the intrinsic capacitance associated with the high resistance of the Faraday cup.

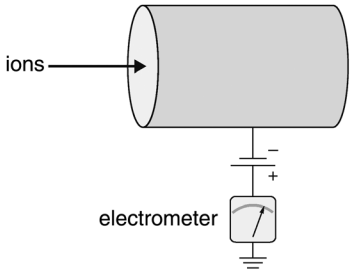


Figure 15.28 Schematic design and operating principle of a Faraday cup.

15.4 Fragmentation Techniques

The fragmentation of ions in the mass spectrometer and the subsequent detection of the produced fragment ions can give a lot of information beyond quantitation. For small molecules the fragment spectra, also called MS/MS or tandem-MS spectra, can give information about the analyzed structure when compared with database information. In biological mass spectrometry fragmentation is mainly important for the identification of peptide sequences, but is also useful for the clarification of complex structures of carbohydrates or modified oligonucleotides.

15.4.1 Collision Induced Dissociation (CID)

To obtain information about the molecular ion structure, the parent ion is fragmented and the resulting fragment masses are determined. A frequently used way to fragment molecular ions is by collision with an inert gas. Here, special collision cells are used, through which collision gas passes or the residual gas in the analyzer is used. The latter is the case in CID experiments in electrical or magnetic ion traps. Collision cells are used in triple-quadrupole-, QTOF, and TOF-TOF analyzers. In all cases the gas pressure can be enhanced by opening a valve. In collision cells the pressure is often significantly increased locally within the cell by passing a small amount of gas for a limited time to avoid electrical discharges. An example is a TOF–TOF analyzer.

The fragmentation due to high energy collision is named CID (collision-induced dissociation) or CAD (collision-activated dissociation), the fragment ions are named CID ions and the spectra are known as CID spectra. CID experiments in TOF-TOF analyzers are carried out with high kinetic energy, for example, 1000 eV (high-energy CID). Under these conditions a huge

amount of energy is transferred, which leads directly to bond breaks (Figure 15.23). The ions in a quadrupole analyzer have low kinetic energy (of the order of 1 eV). For an effective fragmentation the ions are accelerated by additional collision activation with about 20–150 V. Under these conditions a single collision is mostly efficient enough to fragment analyte ions (low-energy CID; Figures 15.21 and 15.24). Beside the kinetic energy the gas pressure in the collision chamber is an important control parameter. With increasing pressure the amount of multi-collision events and the excitation energy increases. Assessments in such cases reveal that a reduction of signal intensity of a precursor ion to about 30% means that on average each single precursor ion has one collision. In electric and magnetic ion traps *low-energy* CID fragmentation is also carried out. In these cases the residual gas, mostly helium, is used for fragmentation and the speed of ions is, in addition, systematically increased.

First the ions of interest are isolated. In the electric ion trap the ions with small *m/z* values are ejected from the trap through elevation of *q* whereas the ions with large *m/z* values are strongly excited by a frequency mixture at the end caps so that they collide with the walls. This process can be performed very exactly: for example a monoisotopic peptide ion can be selected. Afterwards, by resonant excitation, the isolated ions enter a higher orbit (with a greater distance to the middle of the cell), increase their speed, and collide with helium atoms under *low-energy*-CID conditions. In this respect the inner energy increases due to multiple collisions until the critical point is reached and the fragmentation starts. Whereas the peptides are still captured in the trap, the dissociation can be controlled by variation of the amplitude and the duration of excitation. In many cases the fragmentation efficiency for the precursor ions is 100%. In tandem MS experiments, the daughter ions are then analyzed – they are ejected one by one out of the trap and detected. But in many cases, if the structure of a substance is to be elucidated, the information given by the MS/MS spectra is often not sufficient, because only a few fragment ions occur or because of the complex structure itself. It is possible in such cases to perform a series of fragmentation experiments. Instead of direct analysis of fragment ions, selected fragment ions are isolated and further fragmented. The resultant ions are analyzed (MS^3). Within the scope of so-called MS^n experiments up to ten MS cycles can be performed. In practice more than four cycles are rarely used, because either the amount of available ions is too small or no further information gain is expected. The amount of energy required to induce analyte ion fragmentation depends on the stability, structure, and size of the ion. To achieve optimal results an individual adaption of the CID parameter is necessary. An important difference between *high-* and *low-energy* CID is the timeline of induced fragmentation. In the former it happens directly without any remarkable delay. The consequence is a quite unspecific reaction and therefore a high amount of different fragment ions from the large molecular ions with additional internal fragmentation with multiple bond dissociations. In the latter case, low-energy CID, less energy is transferred and the dissociation of ions is delayed. Consequently, the added energy is distributed over all ion bonds. Fragmentation takes place if the bond energy is lower than the dissociation energy. As consequence, low-energy CID fragmentation is more specific, which means there are fewer different fragment ions than in high-energy CID experiments. Molecular ions with predetermined breaking points, the cleavage of which needs only a small amount of energy, will be favored.

15.4.2 Prompt and Metastable Decay (ISD, PSD)

MALDI is often called a "soft" ionization method, because the ionization of thermally unstable molecules such as intact proteins can be transferred into the gas phase. It was originally proposed that almost exclusively intact molecular ions and only a small amount of fragment ions could be generated during the MALDI process. In the MALDI spectra one noticed first only those fragments as additional signals that were generated directly in the source. This prompt decay is classified as ISD (in source decay). Further investigations showed that some of the produced molecular ions are fragmented later during acceleration in the electric field or in the field free drift. For metastable decay in the field free drift the term PSD (post source decay) was coined. The fragmentation of analyte ions is induced by the uptake of high amounts of energy in the MALDI process.

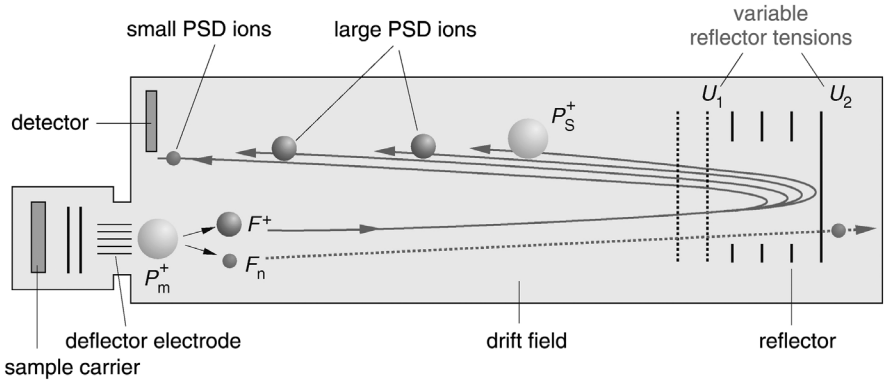


Figure 15.29 Principle of PSD mass analysis in a two-step grid reflector. Molecular ions (*P*+), which decay in the field free drift region, can be detected by stepwise lowering of the reflector potential. By using the deflector electrode it is possible to analyze single molecular ions with a defined *m/z* value out of a mixture.

This can occur due to collision energy within the matrix plume, which is generated directly after the laser hits the sample. The time point after the decay of molecular ions is important for the ability to detect the generated fragment ions in the time-of-flight mass spectrometer. Prompt fragmentation can be observed in the mass spectrum as ion signals with mass corresponding to the fragment ion mass. Fragment ions generated during the acceleration section have different kinetic energies and therefore can contribute to the spectral background noise. PSD ions, which occur in the field free drift, fly with the speed of the precursor ion, and so they cannot be separated by a linear time-of-flight analyzer. By using a reflector time-of-flight analyzer it is possible to separate those ions because, due to the mass loss, the fragment ions have a lower kinetic energy than the intact molecular ion. In the standard operation of a reflector only a fraction of the PSD ions is directed to the detector. Owing to the low energy most PSD ions return too early and do not reach the ion detector. The breakthrough for complete analysis of PSD ions was reached by the development of special reflectors using variable voltage, which could be adapted to the low energy of PSD ions (Figure 15.29). To obtain a complete fragment spectrum (PSD spectrum), a series of part spectra is semi-continuously generated by changing the reflector voltage in a phased manner. Even though this solution is technically outdated it is still occasionally used because no additional expensive components are necessary. Much easier and quicker is the analysis of PSD ions by TOF-TOF mass spectrometers. In this case a collision cell, for example, is not necessary because fragmentation has already taken place. This also means that fragment ions from an active collision cell (collision cell is filled with gas) always consist of a mixture of PSD, CID (high-energy CID), and PSD-CID ions. Consequently, along with the intact molecular ions, PSD ions also undergo collision induced fragmentation. If the decay only results from collision activation, ISD and PSD ions are like CID ions.

Generation of Free Radicals (ECD, HECD, ETD), Section 15.4.4

For MALDI, the contribution of other processes to the "heating" effect of ions is still not clear. It is known that certain matrices such as, for example, 2,5-dihydroxybenzoic acid can induce ionization by fast decay of unpaired electron bonds as well as by collision activation. The efficiency of collision activation depends on the density of the matrix plume after a laser shot and also on the velocity of the analyte ions and matrix molecules. The density of the matrix plume is defined through the laser intensity, the collision energy through the applied acceleration voltage. With the decay of a singly charged molecular ion a fragment ion and a neutral fragment molecule is produced.

ISD ions are not unique to MALDI but can also be induced by ESI. Therefore, the voltage applied to the source is increased, so that the analyte ions are "heated up" due to having a higher velocity during collision with nitrogen molecules and the analytes dissociate. If weak electrical fields are applied, such as, for example, 10 V, the ions move relatively slowl and the collisions result in energy transfer from the ions to the gas molecules. If the ions move quickly a collision induces conversion into kinetic energy and the internal energy of the ions increases. The internal energy defines the average lifespan, if this is not limited by another factor. The bigger the internal energy, the quicker the ions dissociate.

In contrast to MALDI-TOF-MS the combination of MALDI with other mass analyzers plays no analytical role. The same applies to ESI-MS and PSD ions because in both cases the time frame between ion generation and their analysis is too big. Owing to the extended time in the

source, delayed dissociation over several microseconds leads to ISD ions. Such ions could be used by several mass analyzers that do not have a tandem function integrated. In all cases ISD fragment ions will be further analyzed by MS/MS. Such experiments are often named pseudo-MS/MS/MS or pseudo-MS^3.

15.4.3 Photon-Induced Dissociation (PID, IRMPD)

By absorption of photons the internal energy of molecular ions can be increased and this can induce dissociation (photon induced dissociation, PID). In biological mass spectrometry, this approach has so far been implemented routinely only using the magnetic ion trap. For this purpose, the trapped ions are irradiated in the analyzer through a window with an infrared laser. The energy of the photons is comparable to the most frequently used MALDI-UV lasers but is about an order of magnitude lower. This means that the molecule ions absorb many photons before dissociation occurs. Their internal energy is increased in discrete steps. This fragmentation technique is called infrared multiphoton dissociation (IRMPD). The timescale within which a certain number of photons are absorbed depends on the photon density (irradiation strength) and the absorption cross section of the molecular ions. The latter increases with ion size. However, for larger molecular ions the additional energy is distributed over many bonds by vibrations and rotations. This implies that for larger molecular ions a larger number of photons are needed to induce dissociation. The absorption cross-section and the number of photons needed have opposite effects. The IRMPD is a particularly "soft" and easily controlled fragmentation technique, because the required energy for dissociation is supplied in many small, defined steps. The uptake rate can be easily controlled by varying the laser irradiance. In low-energy CID, molecular ions are heated, usually, in several steps (through multiple collisions). In contrast to IRMPD the supplied energy in each step is not equal and may, depending on the type of impact, vary considerably. IRMPD and CID spectra of the same molecule ion often differ in the number and nature of the detected fragment ions and their frequency. Therefore, IRMPD and CID are also combined in MS^n experiments. Here selected IRMPD ions are investigated by CID and vice versa. Because MALDI is a photon-induced process, one may call the resulting fragmentation processes "photon induced". But this is not common, because in MALDI the analyte molecules and not the matrix molecules are the target of the photons. Sometimes IR lasers are used instead of UV lasers. In such cases, typically, the analyte molecules absorb photons and this increases their internal energy. Calculations show that the irradiances used in the IR-MALDI are much too small to fragment the analyte molecules with PID. In such cases, other processes, including CID, are responsible for the fragmentation of the ions.

15.4.4 Generation of Free Radicals (ECD, HECD, ETD)

In chemmistry, molecules with an unpaired electron are referred to as radicals and are very reactive. Mass spectrometry uses this reactivity by, for example, combining multiple positively charged molecular ions with a free electron. The internal energy of large molecular ions is thereby changed only minimally, but their fragmentation occurs spontaneously.

New reactions involving unpaired electrons are responsible for the fragmentations. These reactions lead to the reforming of bonds involving electron pairs to generate stable pairings under the given conditions. The use of unpaired electrons for fragmentation of large biological molecular ions was first established with magnetic ion traps. To this end, the trapped ions are bombarded with electrons using an electron gun. This technique is known fragmentation electron capture dissociation (ECD). By the inclusion of an electron a radical molecular ion with a reduced charge number arises from a multiply charged molecular ion:

$$\left[M+nH^+\right]^{n+} + e^- \rightarrow \left[M+nH^+\right]^{(n-1)+}$$

An important parameter for the performance of ECD experiments is the kinetic energy of the electrons. When the kinetic energy is very low (<0.2 eV) the capture probability is high enough to achieve acceptable reaction yields. When the kinetic energy is a little higher the probability is reduced significantly. If it is 1 eV, for example, the capture probability can be two to three orders of magnitude smaller. Under these conditions ECD is no longer efficient and electron capture is an almost negligible side reaction. Here, inelastic electron–ion collisions dominate,

which increases the internal energy of the ions, similarly to ion–neutral collisions with *low-energy* CID. Systematic studies have shown that the capture probability of electrons as a function of their kinetic energy has two maxima, a relatively narrow one at 0 eV and a much wider one at 10 eV. The trapping of higher energy electrons is connected to prior electronic excitation. This means that the internal energy of the ions is remarkably enhanced. Consequently, in many cases, spontaneous fragmentation occurs and ECD follows by further, secondary fragmentation. This includes bond breaks such as preferred at low- and high-energy CID experiments. This ECD variant is called hot ECD or HECD. HECD spectra of larger molecules are for the reasons given above usually more complex than the corresponding ECD spectra and can therefore enhance the information content. Supplemented by ECD and HECD, a modern high-resolution and fully-equipped FT-ICR mass spectrometer allows very detailed structural analysis of relatively large molecules, particularly medium-sized proteins (<50 kDa), given sufficient quantity (e.g., 1 nmol). In these cases, it is possible to answer remaining structural questions through a combined targeted series of MS^n-experiments: ISD, CID, IRMPD, ECD, and HECD. A second possible way to transfer single electrons to positively charged molecular ions is derived from the chemical ionization of molecules. For this purpose an ionization reagent, usually a small organic molecule, is ionized using an electron gun in the presence of an inert gas to slow down the electrons, which are then combined in the gas phase with the neutral analyte molecules. During the collisions the charge is usually transferred to the larger analyte molecules by electron or proton transfer. Such reactions are also possible with oppositely charged molecular ions instead of neutral particles. These reactions are facilitated by the fact that the participating ions are electrostatically attracted and as the result two charges are neutralized. Whether a proton (proton transfer reaction, PTR) or electron (electron transfer reaction, ETR) is transmitted is primarily dependent on the chemical structure of the used anion and the number of electrons. Stable anions with an even number of electrons are suitable for PTR, while ETR requires stable radical anions with an unpaired electron. Fluoranthene (Figure 15.28) is particularly suitable for this purpose. With PTR, protons of multiply positively charged molecular ions can be abstracted to reduce the charge state.

$$\text{PTR} : [\text{M}+n\text{H}]^{n+} + \text{A}^- \rightarrow [\text{M}+(n-1)\text{H}]^{(n-1)+} + \text{AH}(\text{A}^- \text{ anion})$$

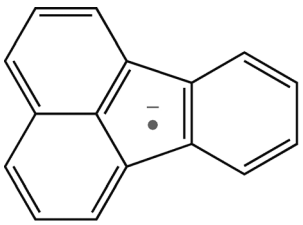


Figure 15.30 Fluoranthene radical anion.

Using ETR, single electrons can be efficiently transferred to large multiply charged positive molecular ions to induce their dissociation. This fragmentation technique is called electron transfer dissociation (ETD):

$$\text{ETD} : [\text{M}+n\text{H}]^{n+} + \text{F}^{-\surd} \rightarrow [\text{F}_1 + n\text{H}]^{(n-1)+} + \text{F}(\text{F}^- \text{ fluoranthene radical anion})$$

An important advantage of ETD versus ECD is that it can easily be integrated in electric ion traps that are available on the market. For this purpose, an additional ion source is connected, in which the radical anions are formed, if necessary. The anions are transferred into the ion trap and then collide with the isolated analyte. A particular advantage of electric ion traps is that ions of opposite charge can be captured easily at the same time in a small volume segment. ETD reactions are very efficient and superior under these circumstances to ECD, because the collision cross section of the charge carrier particles is much bigger in comparison to the electrons of ETD. The integration of ETD in commercial QTOF analyzers is imminent. It will take place in the second quadrupole, which needs to be supplemented and modified technically. The PTR technique was known long before the discovery of ETD and was used to reduce the high charge numbers and the associated large number of different charge states produced by ESI protein cations in MS/MS experiments. It has also been used to reduce the complexity of the recorded mass spectra. This reduces the demands on the resolution and it can also allow the use of cost-efficient, electrical ion traps for these experiments. These analyzers use PTR and ETD in combination, especially for the structural analysis of small and medium-sized proteins (<30 kDa). Because ECD, ETD, and HECD require more positively charged molecular ions, they are almost exclusively combined with ESI. An important observation was that ECD or ETD produces other fragment ions than do CID, PSD, or IRMPD. Another important observation was that a fraction of the ECD or ETD ions, especially larger molecular ions, is only detected when the internal energy is slightly increased, for example, by collisional activation (collisional warming). In these cases, the ECD or ETD fragmentation has taken place without fragment separation. The general rule is that the information content of low-energy CID spectra

Structure Elucidation, Section 15.6.3

decreases with the size of molecular ions. Thus, for medium-sized proteins only a few different fragment ions are detected. This limitation is not exclusive to ECD and ETD spectra. In these cases, ECD or ETD is used first and CID in subsequent experiments to fragment selected fragment ions. These advantages are the reason why ETD in particular has become an important tool in routine biological mass spectrometry.

15.5 Mass Determination

15.5.1 Calculation of Mass

The mass of a molecule with a given elementary composition is calculated in three ways, defined as follows:

- **Average mass** (in the literature often given with an average index): mass of an ion calculated from the average atomic weights of the individual elements with all isotopes (for angiotensin II with the formula $C_{50}H_{72}N_{13}O_{12}$: 1047.21 Da).
- **Monoisotopic mass:** mass of an ion calculated from the exact masses of the most common isotope of an element (for angiotensin II with the formula C_{50} H_{72} N_{13} O_{12}: 1046.54 Da).
- **Nominal mass:** mass of an ion that for simplicity is calculated by using only the integral of the most common mass isotope (this is only valid for small molecules, <1000 Da) (for angiotensin II with the formula C_{50} H_{72} N_{13} O_{12}: 1046 Da).

15.5.2 Influence of Isotopy

Many naturally occurring chemical elements are mixtures of isotopes. This means their nuclei contain the same number of protons (same atomic number) but contain a different number of neutrons. Thus, isotopes of the same element have different masses, but on the chemical level they are almost identical. Consequently, for almost all molecules, different mass signals that reflect the natural isotope distribution are routinely detected. Typically, we observe multiple signals that differ (approximately) by around the mass of a neutron (approximately 1 Da, see below). As a side effect the distances between these isotopic signals can be used to determine the charge of a molecule: for $z = 1, 2, 3$, and so on the distance is 1, 0.5, 0.33, and so on, respectively.

Figure 15.31a shows the *m/z* 400–1000 section of the ESI mass spectrum of a tryptically digested myoglobin (isolated from horse heart) analyzed by an orbital trap mass analyzer with a resolution of 60 000. Figure 15.31b shows a strongly magnified section of *m/z* 689.5–692.5 the ESI mass spectrum with a focus on the isotopic signals of the molecular ions of the tryptic peptide HGTVVLTALGGILK. The total neutron number distribution is determined by the size of the ionized substance, which is reflected by a characteristic pattern of signals and indicates the mass of the sample. For example, the signal difference in *m/z* identifies the charge number $z = 0.5$. The highest signal at 689.93 indicates the lowest *m/z* value. This molecular ion includes only the isotopes ^{1}H, ^{12}C, ^{14}N, and ^{16}O, and thus they have no additional neutron. The barely detectable signal at *m/z* 692.44 correlates with the incorporation of five additional neutrons.

Table 15.2 lists the stable isotopes of the elements H, C, N, O, P, S, Cl, Se, and Br and assigns their exact weight (relative atomic mass) and natural frequency. The increased number of neutrons correlates with the given mass difference. A comparison of these data with the mass of the unbound neutron (Table 15.2, last row) indicates a mass reduction effect during the addition of a neutron. For example, the mass loss during the transition from hydrogen to deuterium is 0.002388 Da while the mass difference in the transition $^{14}N/^{15}N$ is 0.0116 Da.

In general, the actual mass of an atom is always less than the sum of the masses of all the protons, neutrons, and electrons contained in an atom. This phenomenon is called the atomic mass defect. It is the mass equivalent of the binding energy (according to $E = mc^2$) that is released during the assembly of the individual components.

As a consequence of the atomic mass defect even small molecules have many different isotopic variants. This number increases dramatically with the size of the molecules (e.g., proteins). Considering all three natural isotopes of hydrogen, carbon, and oxygen and 18 isotopically different water molecules we find 17 482 different ethanol molecules. For the blood

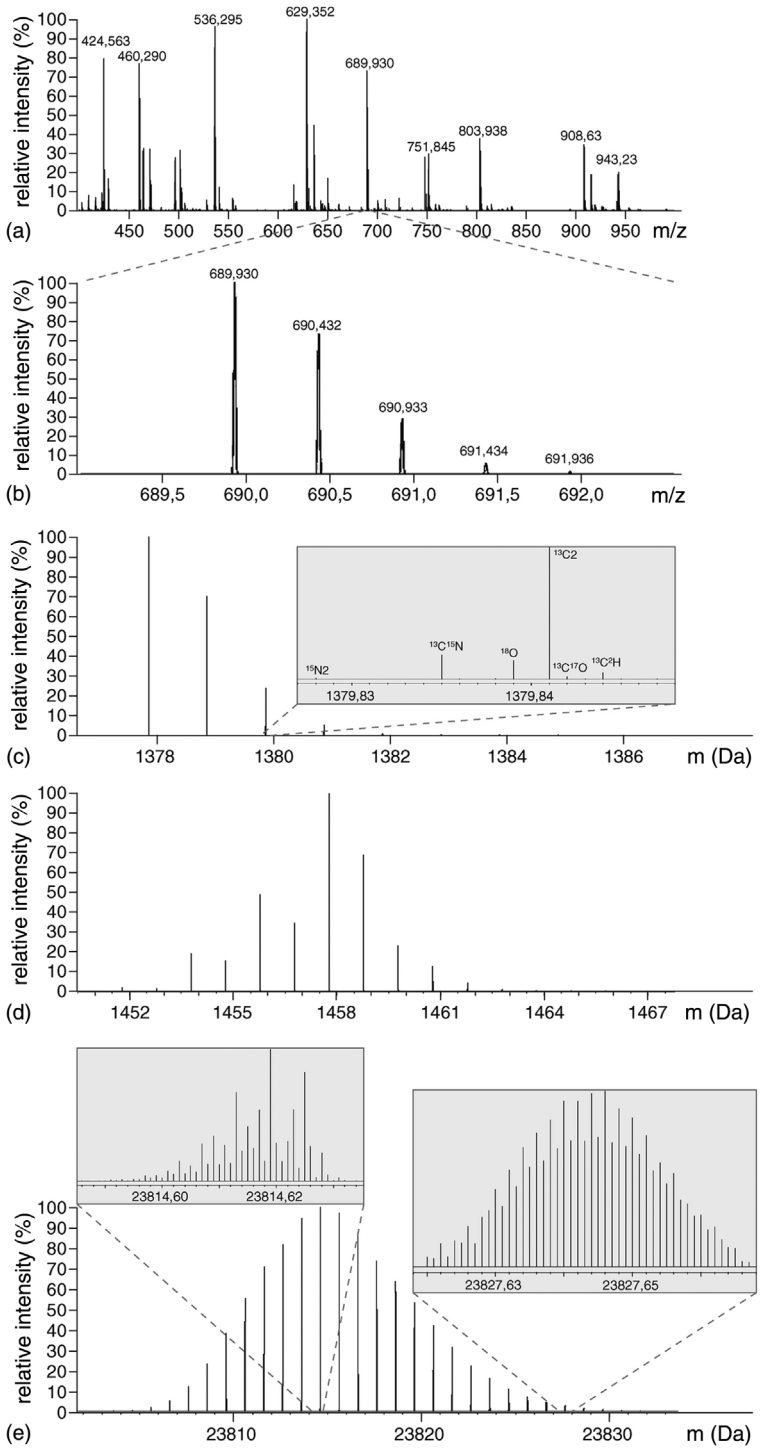


Figure 15.31 ESI mass spectrum of a tryptic digestion of the protein myoglobin isolated from horse heart (a). The *m/z* 689.5–692.5 section showing the isotopic distribution of the doubly protonated molecular ions of the tryptic peptide HGTVVLTALGGILK (b). The calculated isotopic distribution is shown as a line spectrum for HGTVVLTALGGILK (c), for this peptide if the amino acid alanine is replaced with selenocysteine (d), and the human protein complement factor D (e).

pigment heme, $C_{34}H_{32}FeN_4O_4$ (considering the four natural isotopes of iron), there are more than 10^{33} species.

For the following reasons the number of relevant isotopic molecules in MS is much smaller. There is no distinction between isobaric variants with exactly the same mass. However, stereoisomers of the same isotopic composition (isotopomers) can be distinguished after fragmentation of the corresponding molecular ions.

A second reason arises directly from the natural abundance of the involved isotopes. This natural abundance is small or even negligible in many cases. In deuterium it is only 0.0115% and

Table 15.2 Mass and relative abundance of stable isotopes of the elements H, C, N, O, P, S, Cl, Se, and Br. The difference between this value and the one given in the fourth column corresponds to the mass equivalent and binding energy by addition of one neutron. The last line includes the mass of the unbound neutron. Source: Audi, G., Wapstra, A.H., and Thibault, C. (22 December 2003) *Nuclear Physics A*, 729 337–676. http://www.oecd-nea.org/dbdata/data/mass-evals2003/mass.mas03round.

Isotope	Mass (Da)	Uncertainty (Da)	Difference (Da)	Frequency (%)
^{1}H (H)	1.007 825 032 07	0.000 000 000 10		99.99
^{2}H (D)	2.014 101 777 8	0.000 000 000 4	1.006 277	0.01
^{12}C	12.000 000 000	0.0		98.93
^{13}C	13.003 354 837 8	0.000 000 001 0	1.003 355	1.07
^{14}N	14.003 074 004 8	0.000 000 000 6		99.63
^{15}N	15.000 108 898 2	0.000 000 000 7	0.997 035	0.37
^{16}O	15.994 914 619 56	0.000 000 000 16		99.76
^{17}O	16.999 131 70	0.000 000 12	1.004 217	0.04
^{18}O	17.999 161 00	0.000 000 70	1.000 029	0.20
^{31}P	30.973 761 63	0.000 000 20		100
^{32}S	31.972 071 00	0.000 000 15		94.93
^{33}S	32.971 458 76	0.000 000 15	0.999 388	0.76
^{34}S	33.967 866 90	0.000 000 12	0.996 408	4.29
^{36}S	35.967 080 76	0.000 000 20	1.999 214	0.02
^{35}Cl	34.968 852 68	0.000 000 04		75.78
^{37}Cl	36.965 902 59	0.000 000 05	1.997 050	24.22
^{74}Se	73.922 476 4	0.000 001 8		0.89
^{76}Se	75.919 213 6	0.000 001 8	1.996 737	9.37
^{77}Se	76.919 914 0	0.000 001 8	1.000 700	7.63
^{78}Se	77.917 309 1	0.000 001 8	0.997 395	23.77
^{80}Se	79.916 521 3	0.000 002 1	1.999 212	49.61
^{82}Se	81.916 699 4	0.000 002 2	2.000 178	8.73
^{79}Br	78.918 337 1	0.000 002 2		50.69
^{81}Br	80.916 290 6	0.000 002 1	1.997 954	49.31
Neutron	1.008 664 915 7	0.000 000 000 6		

in tritium it is only a trillionth% compared with the 99.9885% of atomic hydrogen. This means natural tritium-containing molecules cannot be detected at all while deuterium-containing variants are only detected if the molecules contribute more than 100 hydrogen atoms. In the case of carbon it is different. This can be explained by the natural abundances of its isotopes, namely, ^{12}C (98.93%), ^{13}C (1.07%), and radioactive ^{14}C (ca. 0.000 000 000 1%). This means that ^{13}C can be detected in even small molecules while ^{14}C cannot be measured. The elements N and O show intermediate behavior (^{15}N: 0.37%, ^{18}O: 0.20%).

Comparing the total mass increase, the effect of the addition of one neutron and thus the effect of the mass defect is small (Table 15.2).

Regarding MS application we need to consider not only the number of detectable isotopic variations of a substance and their frequency but also the resolution of the mass analyzer used.

The isotopic distribution (frequency versus mass) of molecules can be calculated from the corresponding binominal distribution:

$$\left(I_{X_C} + I_{Y_C} + I_{Z_C} + \ldots\right)^{n_C} \left(I_{X_H} + I_{Y_H} + I_{Z_H} + \ldots\right)^{n_H} \left(I_{X_O} + I_{Y_O} + I_{Z_O} + \ldots\right)^{n_O} \quad (15.15)$$

where: C, H, O are elements,
X, Y, Z are isotope masses,
I_{X_C}, I_{Y_C}, I_{Z_C} are the isotope frequencies of element C,
n_C, n_H, n_O, is the total number of atoms of an element in one molecule.

After calculating the cumulative distribution it is normalized to the maximum intensity.

Figure 15.31c compares the measured (Figure 15.31b) and the calculated isotopic distribution for the peptide HGTVVLTALGGILK. The addition of two neutrons is shown: the mass differences result in the observation that all signals (except the first) have a fine structure. This arises from the different mass differences that occur when an H, C, N, or O receives an additional neutron. To fully resolve this fine structure a resolving power of several million would be needed. This is at least partly possible by using a modern magnetic ion trap and a small molecule. For highly accurate mass determination this has important consequences. If the fine structure is not resolved (except the first signal) the present signals reflect the mean mass of a sub-group of the isotopic molecules. To calculate them exactly the sum formula and frequency of the isotopes must be considered. For the first signal the mass of the molecules is equal to the sum of the masses of all the atoms involved. This is because these molecules contain no additional neutron.

Figure 15.31e shows the calculated isotopic distribution for the human protein complement factor D with the formula $H_{1648}C_{1047}N_{309}O_{306}S_{10}$. The frequency of the variant, which is composed solely of ^{1}H, ^{12}C, ^{14}N, ^{16}O, and ^{32}S is only 0.08%. This means that the corresponding ions cannot be detected due to their small amount. Instead a 20 Da wide distribution of isotopes, or instead signal that involves this, is visible. In any case, the available ion current is distributed to many species, which reduces the detection sensitivity. This is a general problem in the analysis of large molecules by MS. The two enlargements in Figure 15.31e show how the complexity of the signal fine structure increases with increasing number of additional neutrons.

The elements P and S are very abundant components of biological molecules. Phosphorus has only one natural isotope (^{31}P) and is therefore not critical for the MS of large molecules. Sulfur, though, occupies a special position. It has four natural isotopes (^{32}S, ^{33}S, ^{34}S, ^{36}S), which means that sulfur atoms affect the characteristic isotope pattern of biological molecules. The same effect is seen with the less frequently observed components Se, Cl, and Br. For bromine the content of the natural isotopes ^{79}Br and ^{81}Br is nearly equal. In selenium the first isotope ^{74}Se has a frequency of 0.89% while the fifth (^{80}Se) has one of 49.61%, which makes it the most common.

Figure 15.31d shows how the calculated isotopic distribution for the cleavage peptide HGTVVLTALGGILK changes when alanine is replaced by the 21st proteinogenic amino acid selenocysteine. Instead of carbon the isotope distribution is now primarily impacted by selenium. This can cause problems, because the automated recognition of peptide signal patterns and the accompanying mass determination are often based on the carbon distribution pattern (Figure 15.31c). The pattern shown in Figure 15.31d could be misinterpreted as a superposition of the signal patterns of four different peptides.

15.5.3 Calibration

An important prerequisite for mass determination is the calibration of the mass spectrometer. For this purpose appropriate calibration substances are used with a known exact mass. Depending on the quality of the device and its sensitivity to temperature fluctuations the instrument must be calibrated once a year, weekly, daily, or before and during each measurement. The best results are obtained when each additional recorded mass spectrum is calibrated internally. This requires that the sample contains at least one known substance and is detected without interference. Thus, the corresponding signal is not superimposed, for example, by another signal. Internal calibration at the end of the evaluation can take place before all other signals are analyzed. The result is a correction of each spectrum based on the reference signals.

15.5.4 Determination of the Number of Charges

When isotope distributions are resolved, the number of charges can, as described above, be derived directly from the space between signals (1, 0.5, 0.33, etc., for $z = 1$, 2.3, etc.). For large molecules the isotopic distribution is often not resolved and also the width of the peaks does not allow unambiguous information on the corresponding charge state. In such cases it is often possible to identify successive ion signals (e.g., 5+, 6+, 7+) that belong together. If this is

successful, the number of charges n and thus the average molecular mass from the measured m/z-ratios m of any two consecutive molecular ions ($m_2 > m_1$) can be calculated as follows:

$$m_1 = \frac{M + nX}{n}(1) \quad \text{and} \quad m_2 = \frac{M + (n-1)X}{(n-1)}(2) \tag{15.16}$$

where X is the mass of the charge carrier. For positively charged, protonated molecular ions $[M + nH]n$, this is the mass of the proton (1.007 Da). For negatively charged, deprotonated molecular ions $[M - nH]^{n-}$ the algebraic sign changes. From (1) and (2) in Equation 15.16, n can be calculated as follows:

$$n = \frac{m_2 - X}{m_2 - m_1} \tag{15.17}$$

and:

$$M = n(m_1 - X) \tag{15.18}$$

15.5.5 Signal Processing and Analysis

To evaluate the recorded mass spectra, specific programs for the ionization method (MALDI or ESI) and the mass analyzer are used. Depending on the nature and quality of the raw data, these are also preprocessed before the mass determination. Current methods are the withdrawal or correction of the background signal (baseline) and the removal or reduction of signal noise by using appropriate filters, which may be electronic or chemical. In the next step signals are selected and assigned to a m/z value. Typical criteria include the requirement of a minimum signal-to-noise ratio (e.g., 6 or 10) and a maximum peak width at a particular height (e.g., m/z of 0.2 at half-height). Different methods are used to determine the m/z values. These take as a basis the highest point of the peak or, most commonly, the centroid (center of gravity mass) of the total peak or a defined upper part. Such a limitation is common practice and takes into consideration known problems such as, for example, that the peaks are asymmetrically distorted in the lower part or exhibit other artifacts.

Other, more elaborate methods calculate an expected isotope pattern for a certain class of substances (e.g., peptides or oligonucleotides) and calculate an expected isotope pattern for a particular m/z range and compare this with the observed patterns. This adjustment often also includes, in addition to the determination of m/z, additional acceptance criteria. When the signal series shows an atypical pattern for a peptide, these peaks can be excluded from the determination of mass or can be evaluated accordingly.

15.5.6 Derivation of the Mass

In the last step the m/z values are used to determine the molecular weights. For singly charged molecular ions, which are preferably formed by MALDI, only the mass of the charge carrier must be subtracted from the m/z values. For negatively charged molecular ions this amount will be added accordingly. For multiply charged molecular ions the m/z values are multiplied by the respective number of charges and the mass of the charge carriers (e.g., two or three protons) is subtracted or added. For molecules with a molecular weight of up to 5 kDa modern machines can readily resolve the isotope distribution to baseline. In such cases the monoisotopic mass is determined routinely. If the isotope distribution is not resolved, the average mass is determined. This is usually less accurate than the monoisotopic mass. If the isotopic distribution of molecules >5 kDa is well resolved, a more accurate mass determination is possible, but is complicated because the monoisotopic signal is no longer detectable. Among the remaining signals a complex fine structure, and not a single ion species, is hidden. Furthermore, it is not immediately clear to which signal group a given signal belongs, that is, how many additional neutrons are present.

15.5.7 Problems

There are several different causes of false systematic mass determinations. For example, prompt decay of unstable molecules may lead to signals in the ion source (e.g., from

fragment ions), which are wrongly evaluated. Another frequently observed change of the mass of molecular ions, seen with peptides, proteins, and nucleic acids, is caused by exchange of the acidic protons with metal cations, particularly the alkali metal cations Na^+ and K^+. This is important for example for the amino acids aspartate and glutamate and phosphodiester linkages in nucleic acids. For small and medium-sized molecules this exchange causes, if it is not quantitative, characteristic satellite signals (H/Na: +22 Da, H/K: +38 Da). When it is not possible to recognize the replacement of a proton with a metal cation, a protonated molecular ion is assumed and a large mass determined. When the molecules are huge, their isotopic distribution is not resolved, the signal is often so wide that the satellite signals are also not resolved. Depending on the ratio and resolution the peak shift results in a peak broadening or formation of a "shoulder." Such effects can also be caused by the addition of sample buffer components or matrix molecules (MALDI) (adduct formation). Possible cleavage reactions, in particular the loss of a water or ammonia molecule, act in the opposite direction. Other reasons for signal loss are satellite signals or signal distortions that are unwanted derivatization reactions during the sample preparation. An example of this is the oxidation of methionine ($+O_2$: +32 Da) and tryptophan ($+O_2$: +32 Da) by traces of ozone.

Other examples relate to undesired side reactions with the components of buffer solutions or separation media (Figure 15.32). A typical example is the addition reactions of the sulfhydryl groups of cysteine side chains with the reducing agent mercaptoethanol or non-polymerized acrylamide monomer during gel electrophoresis. In addition, the free amino groups of the N-terminus and from lysine side chains are reactive and can, for example, be modified during a BrCN cleavage by formic acid to formyl derivatives.

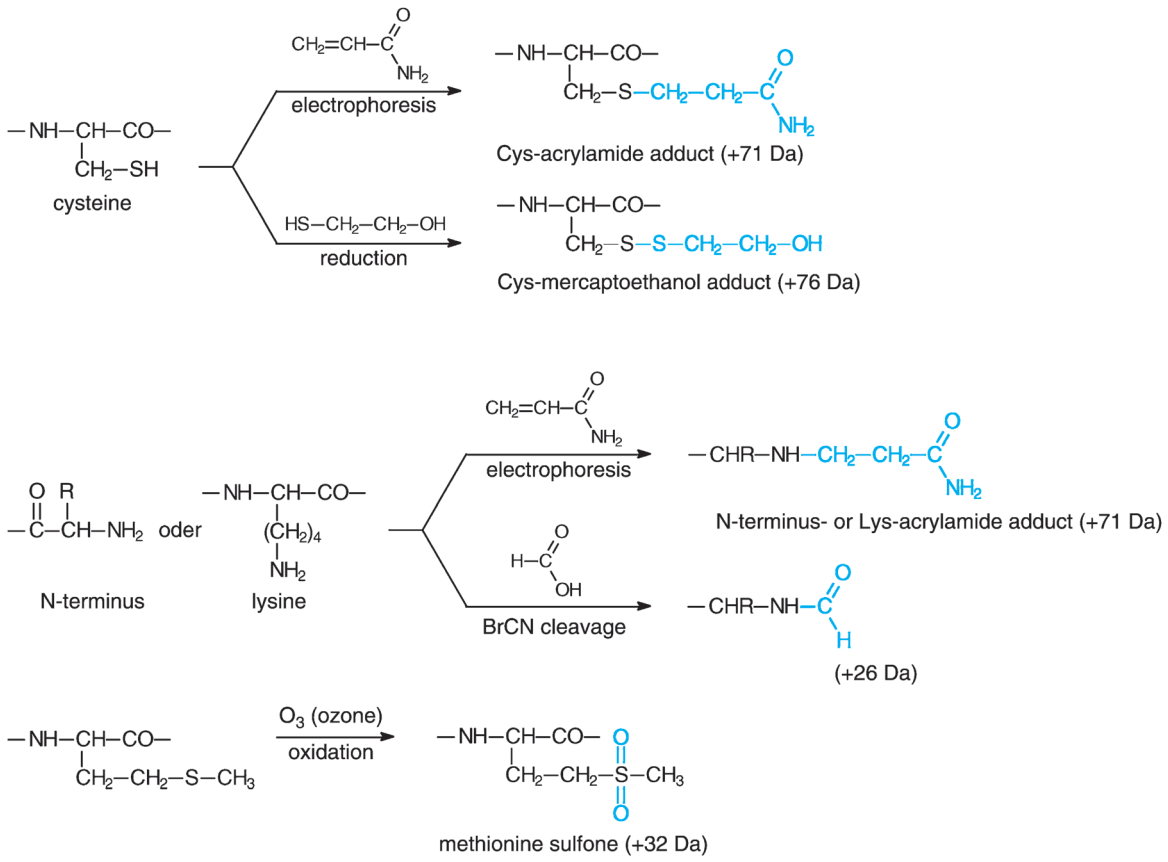


Figure 15.32 Examples of protein-modification by buffer solutions or separation media.

15.6 Identification, Detection, and Structure Elucidation

In practice, attempts are made to answer different questions by mass spectrometry. These can be summarized under the terms identification, detection, and structure elucidation. Here, identification means an answer to the question: "Which molecule (from a number of known molecules) is in this sample?" A proof would be the answer to the question: "Is molecule X present in this sample?" With structure elucidation it is also possible to answer the following question for unknown molecules: "What kind of molecule is this?" The next section presents the most important techniques and methods with which biological molecules can be identified by mass spectrometry. Section 15.6.3 also describes unknown structures that can be elucidated by mass spectrometry.

15.6.1 Identification

The mass of every molecule is defined by its molecular formula of elements. Based on this knowledge it can in some cases be sufficient to calculate the intact mass for identification. However, when different structures have the same molecular formula (isobaric molecules) the information about their mass is not sufficient for identification. In such cases further information for identification is needed. Using MS/MS experiments it is possible to gather such information. When analyzing large biological molecules several other factors make an identification based solely on molecular mass impossible. Proteins are a very good example to illustrate this problem. The molecular mass of a protein is defined by its amino acid sequence but there are also a large variety of modifications like post-translational modifications, polymorphisms, or chemical modifications. These modifications can lead to a significant mass deviation between the known amino acid sequence and the functional protein. When the protein is identified using another technique a difference in mass is a clear indication of one or more modifications of the primary structure.

To identify substances using mass spectrometry several methods are well established. The method of choice in most cases reduces the number of candidate molecules by filtering for a distinctive mass. In the next step the masses of fragment ions that were detected by MS/MS analysis are compared to those of the candidates based on their calculated values. The comparison of the acquired data is commonly fully automated and uses computer optimized algorithms and databases that contain as much structural information of the described molecules as possible. Alternatively, spectra libraries can be used, which means that instead of potentially possible ("predicted") fragment ions only those that were identified in other experiments are used for identification. This strategy is widely used for the identification of small molecules like metabolites, for peptides and proteins almost every identification is based on sequence databases. A very important criterion for evaluation of the identification is the reliability and credibility of the data. To satisfy these requirements evaluation parameters based on statistical calculations and probabilities need to be met.

Analyzing very large molecules by the method described above does not always result in satisfying results because either the number of possible fragment ions is too high and their specification to low or, as in most cases, the efficiency of fragmentation is unsatisfactory. To solve this problem the fragmentation in the mass spectrometer is replaced or supported by a specific chemical or enzymatic reaction. During the catalytic reactions very specific but rarely common bonds are cleaved so that large molecules are segmented into a manageable number of small fragments.

Cleavage of Proteins, Chapter 9

The masses of such fragments are detected and compared to the calculated masses of possible candidates. For independent identification, further fragmentation of the selected fragments is possible using MS. The performance of a method for identification can be estimated based on parameters for specificity and sensitivity. The method of choice should identify the correct molecules based on a high probability and the amount of material necessary for analysis should be very low. The specificity of mass spectrometry based approaches is very high especially when MS and MS/MS analyses are combined. The value of specificity rises as the accuracy of mass estimation rises. The smaller the error tolerance, the lower is the number of possible

structures. The efficiency of fragmentation and selectivity during MS/MS analysis influences the specificity of the method.

The sensitivity of a mass spectrometric method varies based on the substance that is analyzed and the purity of the sample. Compared to Edman sequencing the sensitivity for peptide and protein identification is very high. For this reason a sample amount in the range of several femtomoles is sufficient for peptide and protein identification.

Proteome Analysis, Chapter 39

For example, it is possible to identify more than 1000 proteins using one thousand human cells. DNA or RNA samples behave in the opposite way. Other techniques besides mass spectrometry succeed with as little as several attomoles whereas for an identification using mass spectrometry at least 1 pmol of DNA is needed. Although high amounts of DNA can be produced by PCR this limitation is the reason why mass spectrometry could not be established successfully in the field of DNA and RNA molecule analysis.

15.6.2 Verification

In contrast to identification, for verification the formula or structure of the substance is normally known. Therefore, the exact molecular mass which is needed for mass spectrometric analysis and, additionally, expected fragment ions can be derived.

The type and relative abundance of fragment ions of a specific analyte can be determined experimentally. In verification studies such information, or some of it, is searched systematically or used as filter. The above-described strategies for identification of analytes can be also used for verification, with the simplification that no database or spectra library is needed. The highest sensitivity for verification of one or more substances in a sample can be reached with a triple quadrupole mass spectrometer with the help of SRM or MRM (Section 15.2.6). In this case only ions with a specific *m/z* value pass Q1 and are fragmented in Q2; only specific fragment ions with specific *m/z* values can pass the detector. The verification of molecular masses and a set of fragmentation ions takes place on the hardware level as the ion current is used in an optimal way. Sometimes post-translational modifications such as, for example, methylation, acetylation, or phosphorylation can be detected for the specific substance in the sample. This is possible due to verification of modification specific fragment ions or neutral losses. Again triple-quadruple analyzers are a good choice. With the precursor ion analysis the specific fragment ions can be detected because of known cleavage products (e.g., $-H_3PO_4$: – 98 Da). However, special enrichment procedures are needed for such analyses.

Hybrid Instruments, Section 15.2.6

15.6.3 Structure Elucidation

If a substance is not known, the chemical structure can be clarified. Proteins and nucleic acids are regarded as known if the sequence is available. For insight into higher ordered special structures mass spectrometric methods have been used in a few cases. The description of these methods is beyond the scope of this chapter. The following remarks are therefore focused on primary structures and modifications of proteins.

The structure of various small biological molecules, such as amino acids, nucleic acids, metabolites and so on, can be solved by mass spectrometry. For peptides, proteins, and especially genomic DNA and also RNA molecules this is, apart from a few exceptions, not true. For nucleic acids very powerful sequencing systems exist and by far the most well-known protein sequences are derived from DNA sequences. For insight into primary structures or peptides and proteins mass spectrometry is nevertheless a very important technique, because a lot of biologically active peptides, especially toxic ones, and most proteins after translation are modified in a comprehensive manner. Neither the chemical structure of those changes nor the amino acid residues that are affected by them derive from DNA – or protein – sequences.

The *de novo-sequencing* of proteins using mass spectrometry is particularly needed if the genome of the organism is not yet sequenced. Peptides and proteins are, with a few exceptions (e.g., circular peptides), linear molecules. This structure is advantageous for mass spectrometry in two ways. First, single bond breaks meet the requirements to produce structure-specific fragmentations. Second, the usual construction facilitates the data interpretation. If a protein is

fragmented by *low-energy*-CID, ECD, or ETD, one peptide bond is preferred for dissociation, whereas it is a more or less a random event as to which one it is. If several ions from the same peptide are fragmented, a complete fragment ion series can be observed. Two consecutive fragment ions of one series differ from each other by one amino acid residue (Figure 15.33a). This amino acid is identified by the mass difference (mass of free amino acid minus one water molecule). In peptides, six different ion series exist that result from single ion dissociation. They are N-terminal a-, b-, and c-ions and C-terminal x-, y-, and z-ions Figure 15.33b). The C- and z-ions are preferably produced by ECD and ETD, *low-energy*-CID produce b- and y-ions, and PSD and *high-energy*-CID produce all series, but a-, b-, and y-ions are produced more often. Through cleavage of carbon monoxide (CO: −28 Da) a-ions can arise out of b-ions, which explains why a-ions often accompany b-ions but appropriate x-ions are not detected. On using PSD and *high-energy*-CID additional fragment ions arise, which can be traced back to multiple dissociations. These are internal fragment ions, which result from the cleavage of two peptide bonds, or fragment ions that results from the cleavage of one peptide bond and the fragmentation of one or more amino acid residues. The challenge especially in these cases is the correct assignment of the detected fragment ions.

Restrictions arise because leucine and isoleucine are isobaric amino acids and the molecular masses of lysine and glutamine differ by just 0.0434 Da. To distinguish the mass difference of 20 and 21 amino acid long fragments, a relatively exact mass determination below 10 ppm should be performed. This is not a problem with high resolution analyzers in this *m/z* range, but, for example, the available ion traps are not able to do so. Modern Q-TOF and time-of-flight analyzers come close to their limits and fail if the fragments are markedly larger. In both cases (Leu/Ile and Lys/Gln) the amino acid identity can be determined by fragment ions, which show additional bond breakage positions at side chains.

For the interpretation of fragment ion spectra of peptides and proteins based on the correct assignment of fragment ions to the appropriate a-, b-, c- or x-, y-, z-series, some empirical rules can facilitate the assignment. For fragment ions of the a- and b-series a mass shift of 28 Da is characteristic. Other mass differences such as, for example, 18 Da appear after water dissociates from serine, threonine, aspartate, and glutamate, and 17 Da for the dissociation of ammonia from glutamine, asparagine, and arginine. Cleavage of C-terminal peptide bonds from proline is often absent. In contrast, N-terminal peptide bond cleavage is favored, if aspartate is a binding partner. Next to the ion series so-called immonium ions appear in MS/MS spectra, which can be helpful for spectra interpretation. These are internal fragment ions, which consist of only one amino acid residue and correspond to a C-terminal a-ion and a N-terminal y-ion ($RCHNH_2^+$, R = side chain).

MALDI and ESI-MS spectra of peptides in general differ markedly from each other. Using MALDI, almost exclusively single charged peptides are generated out of 10–25 amino acid long peptides. ESI generates mixtures of two-, three-, and also fourfold charged molecular ions that can be detected. Depending on size the two- or threefold charged molecular ion is the most frequent one. Accordingly, molecular ion fragmentation is preferred on singly charged ions for MALDI and on two- or threefold charged molecular ions for ESI sources.

Singly charged peptide ions, regardless of whether generated by MALDI or ESI, lead to cleavage of a peptide bond to produce a fragment ion and a neutral particle. It is therefore possible, and not uncommon, that MS/MS spectra are preferably observed of C- or N-terminal fragment ions.

In peptides, mainly basic residues become protonated, that is, N-terminal amino groups; the side chains of lysine and especially arginine are also preferred locations for the positive charge of the proton, influencing the ion series produced. However, it must be considered that in general a mobile proton (mobile charge) is expected, which changes its location often and contributes actively to the fragmentation reaction. This assumption can be drawn for the generation of b- and y-ions.

Figure 15.34 shows as an example a MALDI-PSD spectrum of a single protonated molecular ion of the angiotensin peptide. The interpretation is listed in tabular form. By the preferred localization of protons on basic amino acids, arginine in position 2 dominates in the peptide fragmentation series of the N-terminal a- and b-series. The characteristic loss of NH_3 is an additional fragmentation of the side chain of arginine. Beside the a- and b-ions a lower frequency of several C-terminal fragment ions is detected (Figure 15.34b).

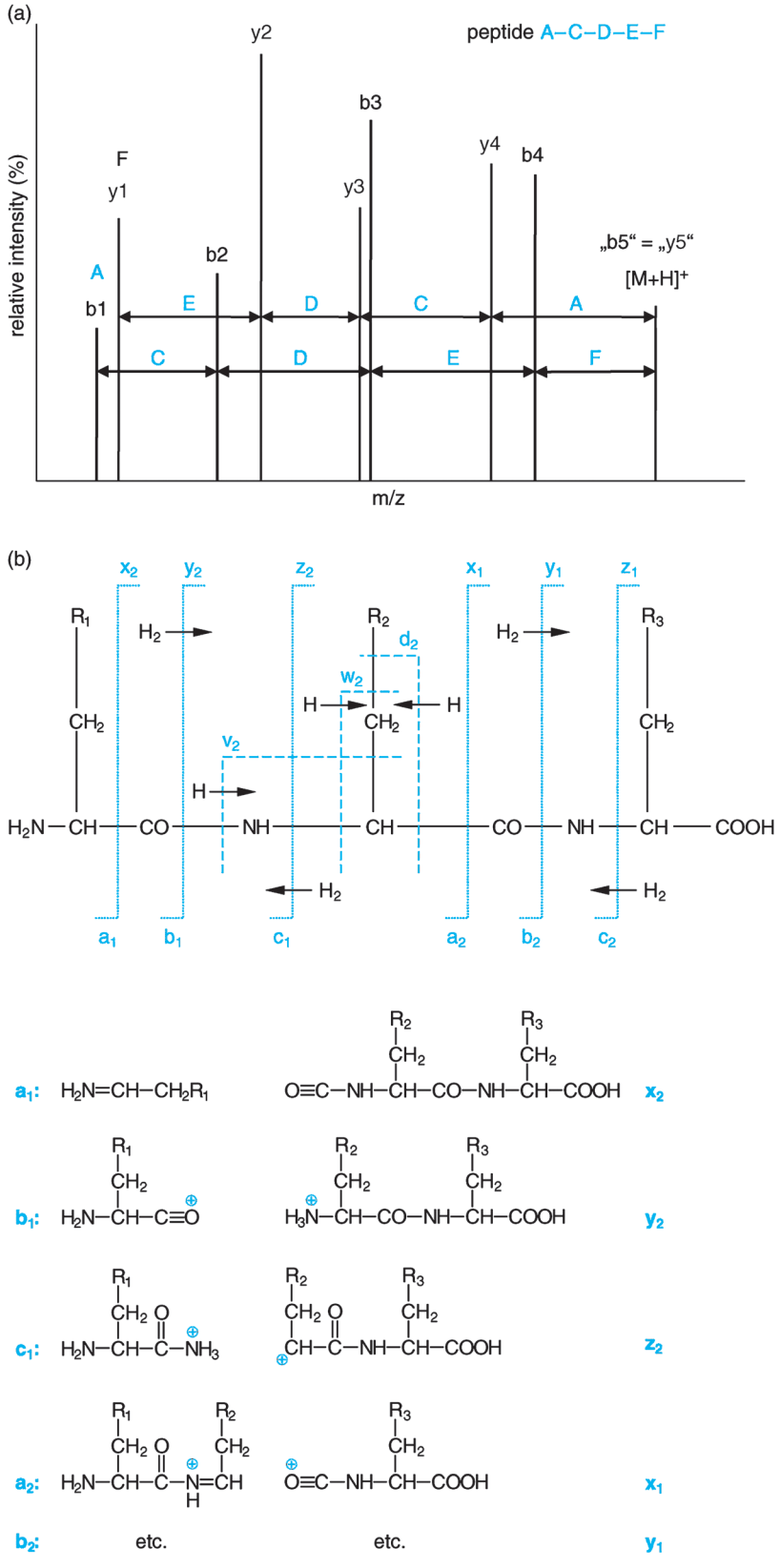


Figure 15.33 (a) Principle of mass spectrometry sequencing of peptides: two consecutive fragment ions of one series differ from each other by one amino acid residue. The one shown is identified by the difference in mass. (b) Nomenclature of the fragmentation of peptides and proteins. Fragment ions that hold the original N-terminus are called a-, b-, and c-ions; C-terminal fragment ions correspond to x-, y-, and z-ions. An additional index gives the amount of amino acids within the fragmention. The y1-fragment ion results from the cleavage of the first peptide bond at the C-terminal side. Correspondingly, the b1-fragment ion is the cleavage product of the first peptide bond at the N-terminal side. The index for the a-, b-, and c-series is therefore in the reverse order to that of the x-, y-, and z-series. Chemically the fragmentation refers to elimination and internal rearrangement reactions. The transferred hydrogen atoms are marked.

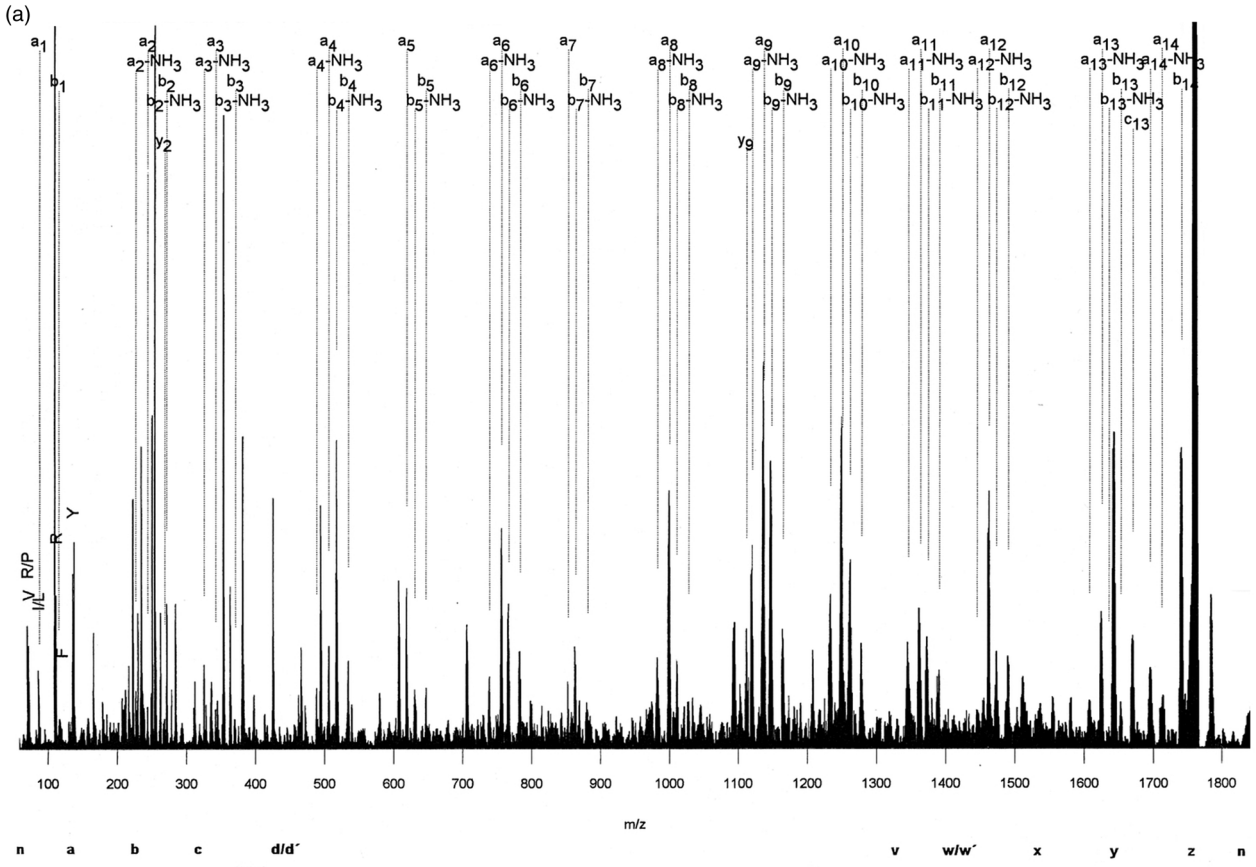


Figure 15.34 (a) PSD-spectrum of angiotensin with the sequence DRVYIHPFHLLVYS and a mass of $MH^+ = 1\,758.9$ Da. Detection of N-terminal fragment ions is mainly observed, because the charge after the decay stays preferable at the arginine. A characteristic loss of NH_3 (b-NH_3-series) is due to an additional fragmentation of the arginine side chain. (b) The interpretation of the data is summarized in the table (figure continued on the next page). Source: Courtesy of B. Spengler and R. Kaufmann, Gießen and Düsseldorf, Germany.

Fragment ion spectra of single and multiple charged peptide ions differ considerably. A noticeable difference is the fact that fragmented ions of multiply charged molecular ions can have larger *m/z* values than the precursor ion so that fragment ions can appear below and above the *m/z* value of the precursor. Additionally the same fragment ions may have different charges, for example, single- and twofold charged b- and y-ions, which means more fragment ions arise in comparison to singly charged molecular ions. Nevertheless a safe and fully automated method to distinguish and assign the isotopes is possible due to different signal distances of differentially charged isotopes (Figure 15.35 below).

Additional differences arise from electrostatic rejection of charges in multiply charged molecular ions.

As consequence the cleavage of peptide bonds results more often in two fragment ions and less often in a fragment ion and a neutral particle. This means that instead of detection of only N- or C-terminal fragments both can be detected and can be used as control for data analysis. A second consequence is that multiply charged molecular ions are more unstable than their singly charged counterparts, which means less excitation energy is needed to fragment them. Along with this, the frequency of internal fragment ions and those that appear from side chains is reduced. As consequence spectra are less complex and easier to interpret. However, it must be considered that the valuable information provided by multiple dissociated ions is lost; such information is needed to differentiate, for example, between isobaric amino acids such as leucine and isoleucine.

In general, the localization of charges and the preferred fragmentation positions depend on the amino acid sequence of peptides. Therefore, a high prevalence of incomplete ion series and signal intensity variations can be observed; consequently, only part of the amino acid

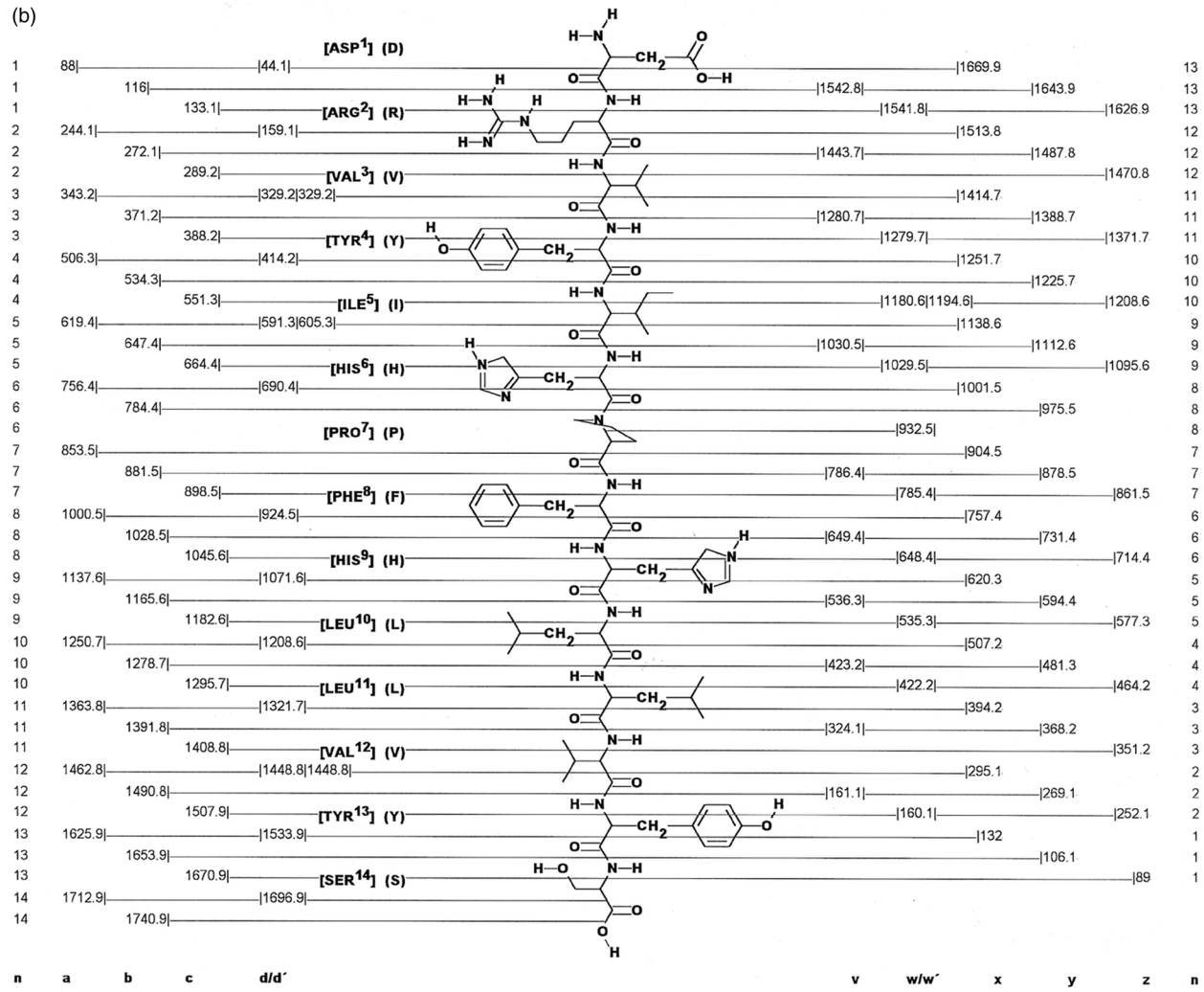


Figure 15.34 (*Continued*)

sequence can be derived reliably. To fill in the remaining gaps, several options exist. One option is to combine *low*- and *high-energy* CID spectra. A second possibility is the combination of CID data with ECD or ETD spectra. Electric and magnetic ion traps provide a third option – MS^n analysis is used to close sequence gaps or to confirm the signal assignment. Therefore, fragment ions, which comprise certain regions, are isolated and further dissociated. Figure 15.35 shows an example. A peptide ESI-MS spectrum was acquired by using an electric ion trap (Figure 15.35a). The signals of singly, doubly, and threefold charged molecular ions are shown. The doubly charged ions were isolated, fragmented by using *low energy*-CID, and the isotope distribution was estimated. The charge was determined from the isotopic distribution (inset in Figure 15.35b).

With the help of the MS/MS-spectra 20 out of 25 amino acids could be aligned. To close the sequence gap, a MS^3-experiment was performed and the b^{11} ion was isolated, further fragmented, and the products analyzed (Figure 15.35c). Because dissociation of N-terminal fragment ions cannot lead to C-terminal fragment ions, and vice versa, MS^3 spectra simplify the interpretation of MS/MS-spectra. In this case b^5–b^{10} could be assigned this way and the internal sequence LKKVAK was derived, meaning that a sequence gap could be closed and additionally other parts of the sequence could be confirmed.

In comparison to peptides the complete structure elucidation of proteins is a highly ambiguous task when using mass spectrometry based methods only. For small proteins the literature gives several examples, for middle-sized proteins (<50 kDa) only some, and nearly

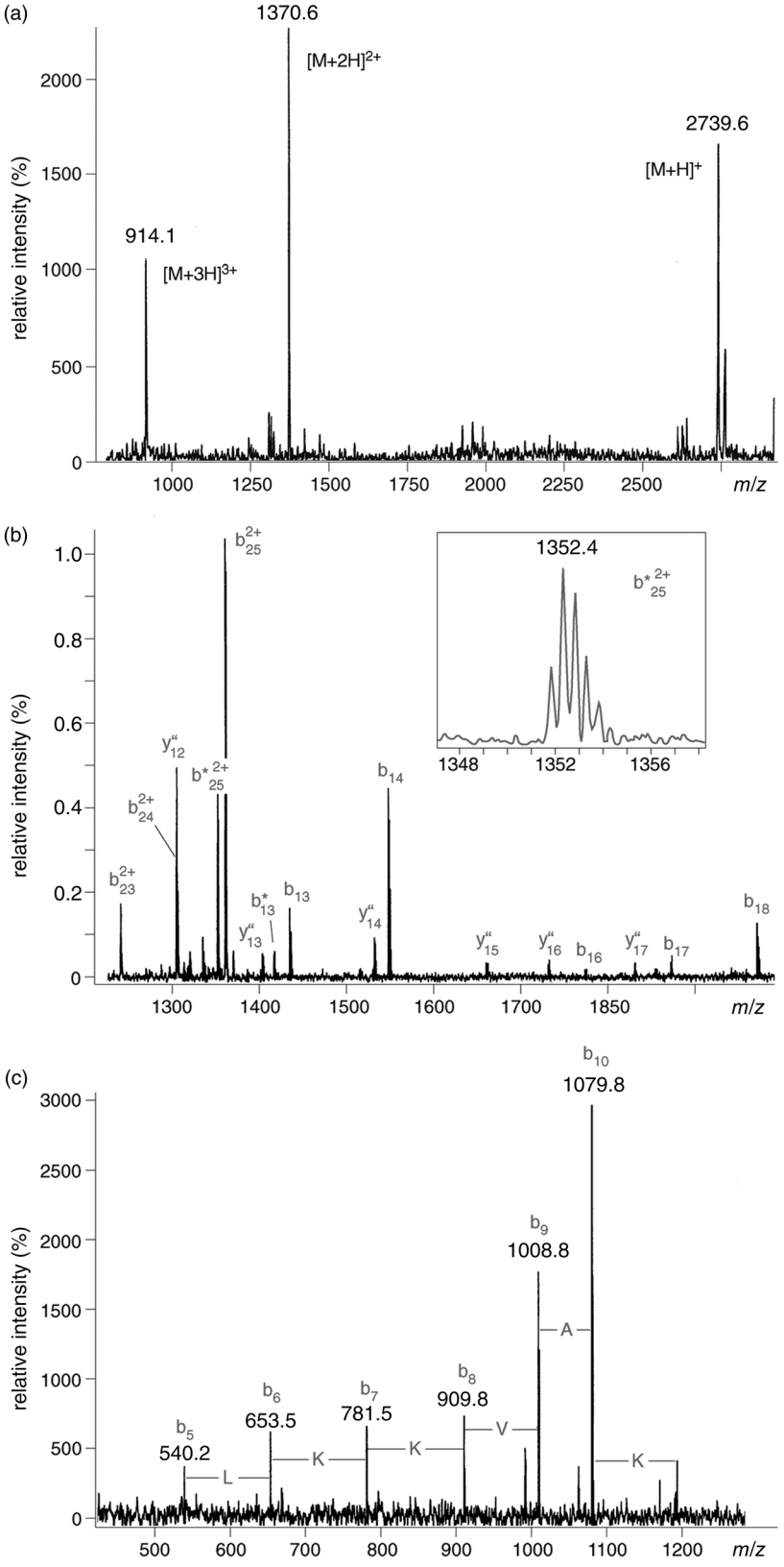


Figure 15.35 MS^n-spectra of a peptide measured with an ESI ion trap. (a) Mass spectrum of the peptide. For sequencing, the doubly charged molecular ion ($m/z = 1370$) was isolated and fragmented. (b) MS/MS-spectrum of m/z 1370 (only the upper m/z range is shown). If the alignment of the signal is not unambiguous or gaps are remaining MS^n-analyses can be very helpful. (c) MS-spectrum of the b_{11} ion (m/z 1208). During fragmentation of b-ions no y-ions can be produced and, thus, newly formed fragments belong to b_5–b_{10} and show the internal sequence LKKVAK. Source: courtesy of Arnd Ingendoh, Bruker Daltonik GmbH, Bremen.

none for large proteins (>50 kDa). In most published cases the structure of the proteins was already known and could be verified or a homologous sequence could be adapted. This is because for larger amino acid sequences there are fewer fragmentation techniques that allow a certain fragmentation of the amino acid sequence. The requirements on resolution and mass accuracy expand enormously along with the number of possible fragment ions and charge states. This results in sequence gaps of Leu and Ile that cannot be differentiated exactly even if different fragmentation techniques and dissociation steps (MS^n) are combined.

Cleavage of Proteins, Chapter 9

An alternative method for the structure elucidation of proteins is the chemical or enzymatically catalyzed cleavage into peptides.

Most commonly used for this approach is the endoprotease trypsin. It catalyzes the specific C-terminal cleavage at the basic amino acids lysine and arginine, if proline is not the neighboring amino acid. For mass spectrometry the specificity of trypsin is a big advantage, because released peptides contain next to the N-terminal amino group, additionally, a second basic group at the C-terminal end. If a complete cleavage of the protein occurred no other basic amino acid is present beside histidine, which is less basic. These are optimal conditions for generating CID spectra of doubly charged peptide ions with long b- and y-ion series, and therefore complete sequence information is available. An attempt to align all cleavage products completely normally does not work. Neighboring peptide sequences in a protein can be recognized by an incomplete cleavage reaction, so that those fragments appear together as one larger peptide [mass (peptide 1) + mass (peptide 2) – mass (H_2O)]. To close such remaining sequence gaps and to rearrange all sequence parts correctly, the protein can be cleaved by a second specific endoprotease, mostly one that cleaves behind acidic amino acids. This digest is also analyzed as described above. Instead of the protein itself different overlapping peptides are sequenced and out of this information a complete sequence is reconstructed. The most efficient way is to combine different approaches, intact protein and cleavage product mass spectrometric analysis, and combine the results.

Beside the sequence analysis of proteins the discovery of chemical modifications, at specific structures and positions of the sequence, is an important component of MS-based protein analysis. Therefore, both described approaches are combined. A MS spectrum of the analyzed protein can often reveal valuable information about the type and extent of the modifications. This refers to the measured deviation of molecular mass and for the unmodified calculated mass and, additionally, possible satellite signals, which rely on incomplete (none quantitative) modifications. If several modifications are present quantitatively, these will be distributed over different peptides if the protein is cleaved. A MS analysis of such products, with comparison of measured and theoretical mass, allows us to localize the possible modification and to exactly determine the inferred mass shift. These values are helpful filters for possible candidates. A phosphorylation, for example, increases the nominal mass by 80 Da ($+HPO_3$), methylation by 14 Da ($+CH_2$), and acetylation by 42 Da ($+CH_2CO$). The goal of the subsequent MS/MS or additional MS^n analysis is to determine the exact localization (which amino acid) and to confirm the probable modification by characteristic fragmentation. If the structure is unknown, MS^n-experiments are very helpful, because they allow the targeted dissociation of fragment ions that contain the modification.

15.7 LC-MS and LC-MS/MS

15.7.1 LC-MS

The advantage of coupling liquid chromatography (LC) and mass spectrometry (MS) is the very efficient separation and concentration of the individual substances and the removal of interfering components such as salts, urea, reducing agents, and buffer substances. Pure samples in the smallest possible volume are the prerequisite for high detection sensitivity, which allows nano-HPLC analysis.

In nano-HPLC flow rates less than 1 µl min^{-1} (e.g., 100 nl min^{-1}) are used and the separation column is coupled online with ESI or offline with MALDI. Offline means that the effluent is fractionated and prepared during the LC run on a MALDI sample carrier The eluate is continuously (for a certain time, e.g. 20 s) transferred to a series of thin-film preparations of the

Chromatographic Separation Methods, Chapter 10

matrix or mixed with a matrix solution at the exit of the capillary. This flow is deposited in defined fractions (e.g., 100 nl). For these reasons normal phase (hydrophilic stationary phase) or RP-LC (hydrophobic stationary phase) are (depending on the substance class) suitable for coupling with MS. Cation and anion exchange LC are not used for direct coupling with MS because the sample molecules are eluted with high salt concentrations and therefore disturb the MS analysis. Many ESI mass spectrometers are used exclusively in combination with RP-HPLC, in particular nano-RP-HPLC. This means the purification, separation, and concentrations are an integral part of the analytical instruments. A key reason for this is that the sample molecules in a volume of, for example, 10 µl can be concentrated (more than 1000-fold) by using nano-RP-LC, which means they are eluted and analyzed in less than 10 nl by online ESI-MS.

In RP-LC the detection sensitivity can be increased by three orders of magnitude (ideal case) while the sample loss is negligible. Another essential improvement for protein and proteolytic peptides analysis is the high separation efficiency of the nano-RP-LC. Thus, the separation is an integral part of the device. In most cases the LC is also controlled by the software of the MS instrument and in some cases even central elements of the LC (column, pre-column switching valve) and the ESI (solid capillary, electrical connections) are integrated in one chip (LC/ESI-chip). In contrast to LC-ESI, LC-MALDI couplings are found only in very specific analytical laboratories. A key reason for this is the offline-coupling.

Modern LC-ESI-MS systems run fully automatically and are highly efficient. They can be used 24 h per day 7 days a week (beside necessary maintenance and service intervals). This efficiency cannot be reached by using a currently available LC-MALDI setup.

Another important benefit of ESI is the significantly better reproducibility of the ion production and therefore the possibility to run LC-MS in a quantitative fashion without the use of stable isotopes.

Other particular advantage of LC-MALDI-MS results from the fact that usually only a negligibly small part of the prepared fractions is consumed in the measurement. This means the prepared fractions can again be measured, stored, and also recovered for additional analyses. In principle, it is possible to integrate additional measurement techniques (light absorption, conductivity, and fluorescence) between the LC and MS step. In most cases a UV detector is used to quantify the elution of the analyte. Figure 15.36 shows an example of the UV chromatogram and the following total ion current (total ion current, TIC) at the detector of the ESI-MS, measured for a proteolytic digest of a protein.

15.7.2 LC-MS/MS

Proteome Analysis, Chapter 39

By coupling nano-HPLC with high-resolution ESI-MS/MS instruments it is possible to identify in 2 h more than 1000 proteins per sample. After proteolytic cleave age the released peptides are separated by reversed-phase chromatography (RP-LC) and ionized online by ESI. Initially, a high level MS-overview spectrum is recorded. Then the most intense peptide signals are determined by software in real time (typically 3–20 for each MS spectrum) and sequentially isolated and fragmented. The change between MS scan and MS/MS scans happens continuously during the LC run. In this case the scanning speed is so high that during a chromatographic elution several cycles are possible (up to 20 MS/MS spectra per second). From the masses of peptides and the recorded MS/MS spectra, the peptides and the corresponding proteins can be identified. With LC-MS/MS mainly ESI is used for the production of ions. For the reasons already mentioned MS/MS-analyzes are often combined with ECD, ETD, and HECD as fragmentation techniques. These techniques cannot be coupled to MALDI because at least doubly charged molecular ions are required for fragmentation. However, the coupling with MALDI offers special advantages that arise from the decoupling of LC, MS, and MS/MS. Here it is possible to take first from all fractions a MS spectrum after the LC run. Based on this information it can be decided which sample molecules will undergo in which fraction an MS/MS analysis. This is in contrast to the online coupling with ESI. There are no conflicts and no time restrictions, for example, overlapping elution of substances with very small mass differences. A new approach for LC-ESI-MS/MS analyses is the so-called multiplex analysis or data-independent analysis. Here there is also a change between survey spectra and fragment spectra. However, no single analyte ion is

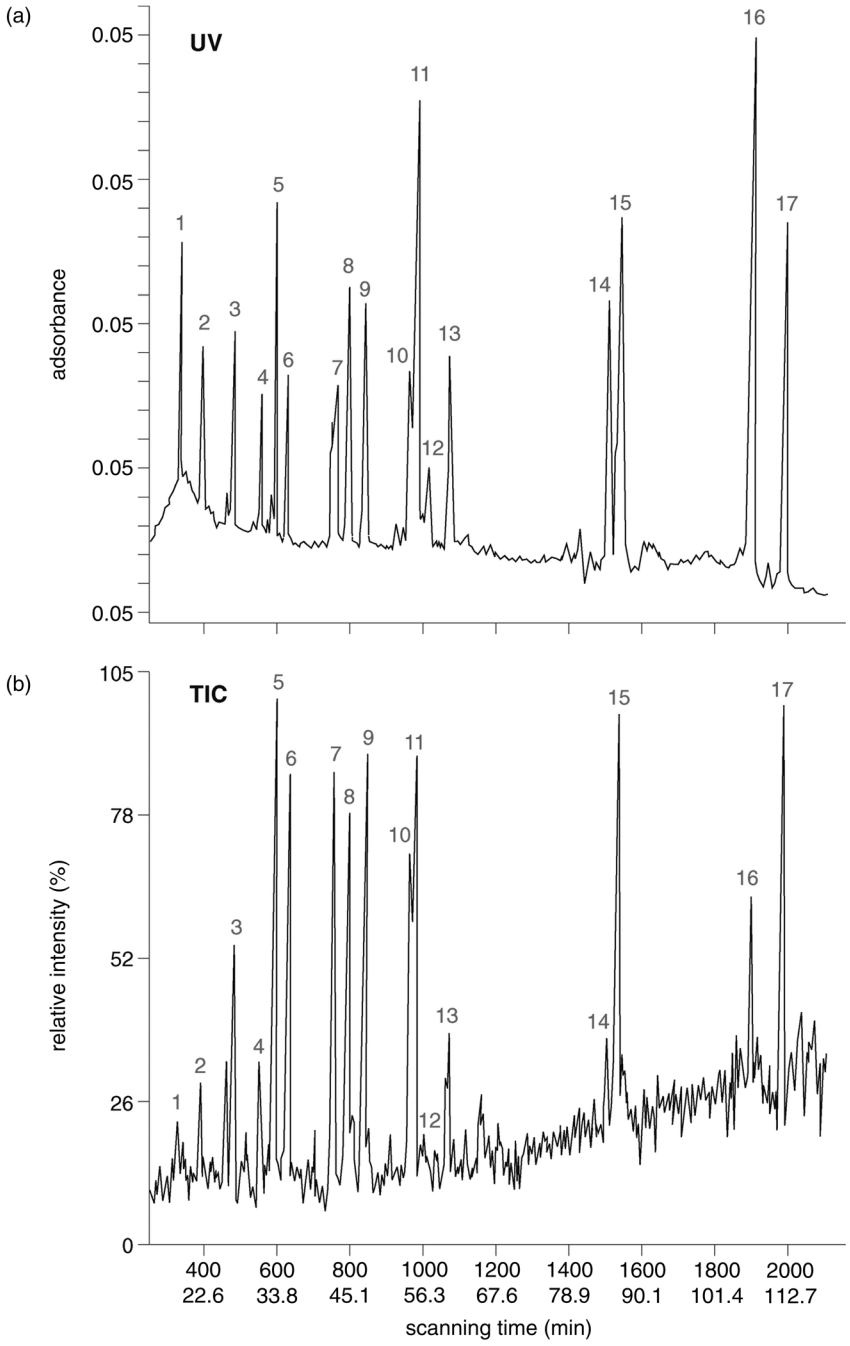


Figure 15.36 Analysis of a proteolytic digest of a protein by RP-LC-ESI-MS. During the LC mass spectra are recorded cyclically. The recording time for each spectrum (scanning speed) is matched to the elution of the chromatographic fractions (in the example shown 3.4 s per mass spectrum). The result of the mass analysis consists of sequentially acquired mass spectra (in the example shown, 2200). Each mass spectrum can be calculated by summing all ion signals to a value that corresponds to the total ion current (TIC). The temporal representations of these TICs give the total ion chromatogram. The individual peaks of the UV signal (a) correspond to the peaks in the TIC (b). Eckerskorn, C. *et al.* (1997) *J. Protein Chem.*, **16**, 349–362.

selected for fragmentation but, instead, all incoming ions are fragmented together with CID and analyzed. To assign the resulting fragments to the correct precursor ion, the chromatographic elution-peaks of all fragments and precursor ions are compared and assigned to the fragments with software support. The simultaneous fragmentation of all molecular ions can be easily and efficiently performed with QTOF analyzers. Two advantages over electrical ion traps are the greater resolving power of the reflector TOF analyzer and the ability to determine the *m/z* values of both molecular ions and all fragment ions to a few ppm accuracy. With increasing complexity of the peptide mixture the method reaches its limits, because the correct assignment of many peptide ions to more fragment ions based on chromatographic elution profiles is not always possible. It could be shown that up to 600–700 proteins (this means thousands of proteolytic peptides) in a sample can be identified securely.

15.7.3 Ion Mobility Spectrometry (IMS)

One method that can be used in combination with LC-MS or LC-MS/MS is ion mobility spectrometry (IMS). Here, ions are moved, using an electric field, through a gas flow that is in opposite direction to their own movement. Based on their drift time through the flight tube, the mobility of the ions can be determined. This depends not only on their mass and charge but also on their shape and size. IMS has long been used for the analysis of small molecules such as explosives or drugs.

In combination with mass spectrometry IMS allows separation of isobaric molecules due to their shape. These may be, for example, peptides that contain the same amino acids in a different order. An analysis of proteins or protein complexes is also possible. IMS can be directly coupled with mass spectrometry and can be found in commercial mass spectrometer applications. Here, IMS is placed between the chromatographic separation and the MS analysis. Typical applications are better allocation of fragments to precursor ions in a multiplex analysis (see previous section) or the study of protein structures and protein complexes.

15.8 Quantification

Mass spectrometry is generally not a quantitative method. The efficiency of ionization of analytes depends on the specific physiochemical characteristics of each sample composition. Comparing the relative frequencies (signal intensities) of different molecule ions allows no conclusion to be made about the concentration in the sample. Moreover, comparison of the same molecule ion in two very different samples is not permitted because the signal can be suppressed by different molecules in the sample. The comparison of an identical molecule in two very similar samples or of chemically nearly identical molecules is valid. Another reason for falsified results is the instability of molecule ions and the corresponding loss of such ions because of possible decay during their journey from the source to the detector (transmission). The ion detector also has an important influence. The detection efficiency of ions with large *m/z* values is decreased compared to ions with a lower *m/z* values. Quantification in the context of mass spectrometry is distinguishable into absolute (absolute amounts or concentrations) and relative quantification (relative variations and changes in percent). Absolute quantification is possible when the sample contains a substance with a known concentration. For relative quantification the addition of a calibration substance is not necessary but needs special care regarding the reproducibility of sample preparation for MS analysis. For the absolute quantification of small molecules the usage of isotopic labeled molecules is well established and allows quantification with very low error values. This technique is based on the knowledge that stable isotopic labeled molecules normally do not differ compared to unlabeled ones. This means that the stable isotopically labeled molecules behave in the same way during sample preparation and show the same ionization and detection characteristics as the unlabeled analytes. There is only one exception, when hydrogen is changed for deuterium the performance of molecules during RP-LC separation changes. ^{2}H is used frequently for heavy isotope labeling. Based on the different masses isotopic molecules can be easily separated from the analytes during detection. This characteristic is commonly used by adding ^{12}C, ^{14}N, or ^{16}O atoms in exchange for ^{13}C, ^{15}N, or ^{18}O atoms, respectively. The crucial issue for the analysis is that the resulting mass shift is large enough so that the isotopic pattern distribution of the "heavy" and "light" labeled molecules overlap slightly. This means that the concentration proportions can be compared in terms of the different signal intensities or the peak area differences. For the analysis of proteolytic peptides the mass difference should be about 4 Da. Absolute quantification is much easier and more precise using stable isotope labeling. Although techniques based on isotope labeling were developed years before MALDI and ESI its application for quantification of large biological molecules like proteins was introduced much later. This was due to the lack of high quality isotope labeled reference substances. However, since a huge variety of methods for the production of stable isotope labeling of different substances has been established there is no longer a problem in purchasing high quality reference substances. Such labeling methods can be divided into chemical, enzymatic, and biological processes. For chemical processes "heavy" and "light" reagent variations are

supplied and can be used to label biological molecules easily. An example of an enzymatic process is an alternative proteolysis of proteins using $H_2{}^{18}O$ or $H_2{}^{16}O$. In this case the C-terminal carboxyl group either binds a "heavy" or "light" oxygen atom. When trypsin was used during the process the exchange of the latter carboxyl-oxygen atom is catalyzed after cleavage, which results in a mass difference of 4 Da.

Proteome Analysis, Chapter 39

For metabolic labeling the growth media of the organism is modified with stable isotope labeled substances. Examples are stable isotope labeled amino acids in cell culture (e.g., human, yeast) or nitrogen sources like ammonium salt for bacterial cultures. One widely used method in this field is called SILAC (stable isotope labeled amino acids in cell culture), which was first published in 2002. Today SILAC is used for various quantitative proteomic approaches.

Further Reading

Ong, S.E., Blagoev, B., and Kratchmarova I. *et al*. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics*, **1**, 376–386.

Ashcroft, A.E. (2003) Protein and peptide identification: the role of mass spectrometry in proteomics. *Nat. Prod. Rep.*, **20**, 202–215.

Han, X., Aslanian, A,. and Yates, J.R. (2008) Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.*, **12**, 483–490.

Watson, J.T. and Sparkman, O.D. (2007) *Introduction to Mass Spectrometry*. 4th edn, John Wiley & Sons Ltd, Chichester.

Steen, H. and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.*, **5**, 699–711.

Mann, M., (2014) Fifteen years of stable isotope labeling by amino acids in cell culture (SILAC), in *Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)* (ed. B. Warscheid), Methods in Molecular Biology, vol. 1188, Springer, pp 1–7.

Protein–Protein Interactions

16

Protein–protein interactions (PPIs) are critical to all biological processes, including DNA replication, transcription, translation, secretion, the cell cycle, and signal transduction. An understanding of these processes requires an understanding of the PPIs involved.

While numerous methods are available for the study of PPIs, a few techniques are predominantly used, such as expression cloning, protein complementation assays (such as the yeast two-hybrid system, Y2H) or affinity purification and mass spectrometry (AP/MS). Other methods are often used to verify interactions resulting from these major methods, for example, FRET (fluorescence resonance energy transfer) may confirm a Y2H interaction.

Each of these methods has their own advantages and disadvantages, so the choice may depend on the specific problem, question, or scale someone is working on. Most PPI methods are more or less qualitative and it remains a major challenge to determine the precise affinity of two interacting proteins. This is particularly relevant as each interaction will require a specific strength for its biological function. For instance, keratin in hair or skin needs to be very stably associated to maintain its mechanical strength. On the other hand, the interaction of a kinase with its protein substrate needs to be transient and quick to permit high turnover and regulation.

Affinity Purification, Section 10.4.8

Immunological Techniques, Chapter 5

Crosslinking, Section 6.3

The wide range of affinities affects their analysis. When proteins or whole complexes are purified they need to survive stringent washing protocols. For example, a purified complex must not disintegrate. However, washing procedures that are too gentle may result in unspecific co-purifications and, thus, unacceptable background noise.

The two-hybrid system detects PPIs in intact cells, usually yeast cells, so that purification is not necessary. Its relatively inexpensive technology has made the Y2H system an often used technique for the analysis of PPIs. The vast majority of PPIs today have been identified by either Y2H or AP/MS. We will thus focus on these two methods while describing several others. Note however, that MS is described in more detail in Chapter 15.

Mass Spectrometry, Chapter 15

16.1 The *Two-Hybrid* System

Peter Uetz and Jitender Mehla

Virginia Commonwealth University, Center for the Study of Biological Complexity, 1000 W Cary St, Richmond, VA 23284, USA

16.1.1 Principle of *Two-Hybrid* Systems

The classical two-hybrid system as devised by Fields and Song (1989) is based on the modular architecture of eukaryotic transcription factors. The yeast Gal4 protein is a typical example and thus has been used since the technique's inception. Gal4 consists of two major domains, a DNA-binding domain (DBD) and a transcriptional activation domain (AD). The DBD binds to

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

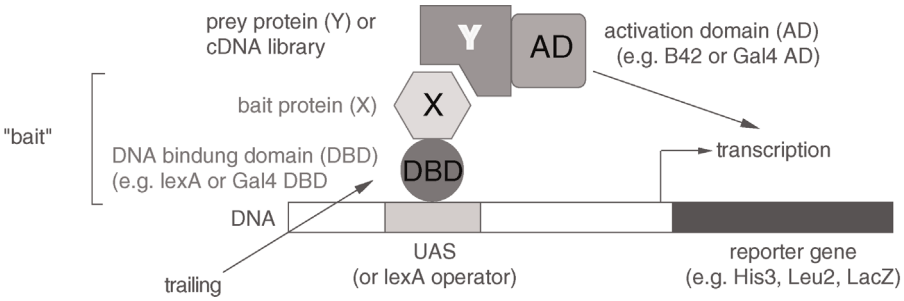


Figure 16.1 Basic concept of the *two-hybrid* system. The functional domains of a transcription factor are separated and fused to two separate but interacting proteins. Transcriptional activation is reconstituted when the two proteins (usually called bait and prey) interact. As a result of the interaction, a reporter gene is transcribed and produces a detectable gene product.

the upstream activating sequence (UAS) of the promoter while the AD recruits the RNA polymerase and so activates transcription. Fields and Song realized that these two domains can be separated and then artificially reunited by other proteins fused to them as long as these other proteins interact. That is, the DBD and the AD do not need to be connected covalently but can be located on two separate proteins that are only linked by a protein–protein interaction (Figure 16.1).

The system requires another component to serve its purpose, namely, a reporter gene with a UAS that binds the DBD and drives transcription. The activity of the reporter gene needs to be detectable, for example, by an enzymatic assay that determines the amount of reporter protein made (see below for details). When bait and prey proteins are expressed separately, no reporter gene activity will be induced because the DBD has no activation domain and the AD cannot bind to the promoter (Figure 16.1).

The Y2H can be used to identify interactions among defined proteins as well as to map interaction domains within proteins. However, one of its most important applications is the screening of whole libraries of "preys" to find new interactors for proteins of interest. To do that the AD is fused to a large set of prey proteins, either as a defined set of proteins or randomly. In the latter case all the proteins that are expressed in a certain tissue may be used. The vast majority of these proteins will not interact with the bait protein but a few may and they will activate the reporter gene and thus can be identified.

16.1.2 Elements of the *Two-Hybrid* System

The original Y2H described by Fields and Song in 1989 used the yeast Gal4 protein and thus was first established in yeast cells. Subsequently, several modifications have been developed that use other proteins or cells, including bacteria and mammalian cells. For instance, Roger Brent and colleagues replaced the Gal4 DBD with the DBD from LexA, a bacterial transcription factor. Essentially any DBD may be used as long as the UAS is also modified. In the case of LexA, the lexA operator sequence was used as UAS. Similarly, a large number of Y2H plasmids to express fusions of bait and prey proteins are now available. A few examples are shown in Figure 16.2.

β-Galactosidase, Section 34.4.3

The lacZ gene from *E. coli* was one of the first reporter genes used. LacZ encodes for the enzyme β-Galactosidase (Figure 16.3). β-Galactosidase converts the indicator compound X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside) into a blue product, so that yeast colonies with an active reporter gene turn blue as long as they are grown on a substrate that contains X-Gal.

Usually additional reporter genes are used to ensure the stability of the plasmids from which the bait and prey proteins are expressed. For instance, bait plasmids may encode the Trp1 gene that allows the cells to grow on media without tryptophan, while prey plasmids may contain a Leu2 gene that allows them to grow on media without leucine (Figure 16.2). By adding additional reporters the stringency of the system can be increased or decreased.

16.1.3 Construction of Bait and Prey Proteins

Once a bait protein has been selected it needs to be cloned into the bait vector. The vectors shown in Figure 16.2 have certain properties to facilitate this:

- A DNA-binding domain, here from Gal4, under the control of the yeast alcohol dehydrogenase (ADH) gene promoter, which is constitutively active. The two bait vectors

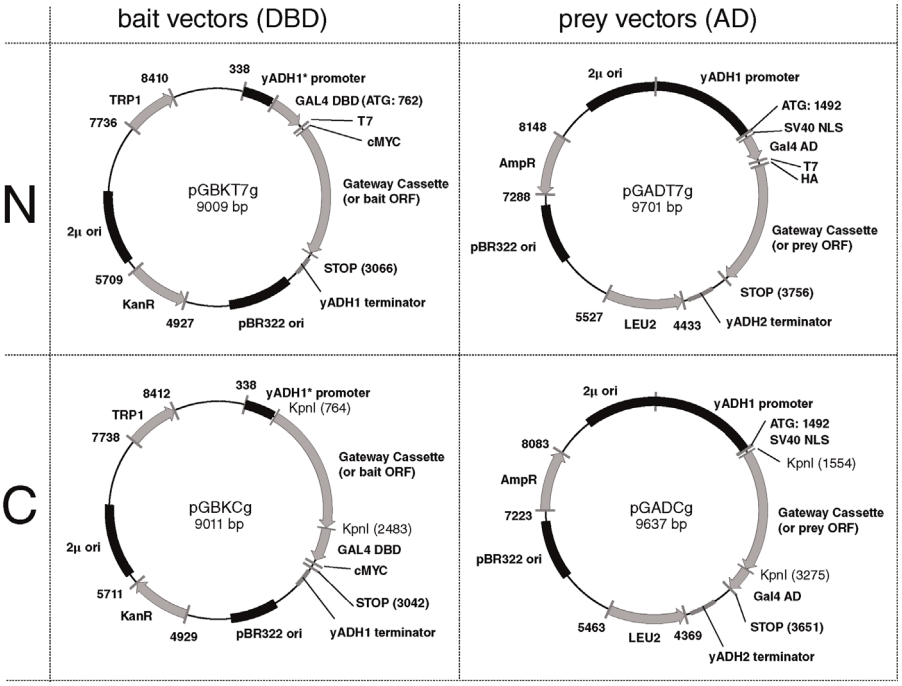


Figure 16.2 Yeast two-hybrid-vectors (expression plasmids). Bait vectors pGBKT7g and pGBKCg and prey vectors pGADT7g and pGADCg. Note that the vectors in the top row produce fusion proteins with the DBD or AD at the N-terminus. The vectors in the bottom row result in fusion proteins with the DBD or AD at the C-terminus of the prey protein. Fusion proteins resulting from the vectors shown can be combined to test the interactions of N- and C-terminal fusion proteins. Source: adapted from Stellberger, T. *et al*. (2010) Proteome Sci. 8, 8.

permit expression as N- or C-terminal fusion of the bait protein. This is relevant as both result in different interactions (see below).

- A cloning site, here a Gateway cassette that permits cloning by recombination (see Chapter 27). Other vectors may have restriction sites that can be opened by a restriction enzyme. The protein needs then to be inserted by ligation (Chapter 27). Note that Gateway recombination sequences are standardized so that the fused proteins are automatically in the right reading frame. This is not guaranteed by restriction and ligation.
- A yeast termination sequence that terminates transcription (here: derived from the ADH1 or 2 gene and, hence, called yADH1 or yADH2 terminator).

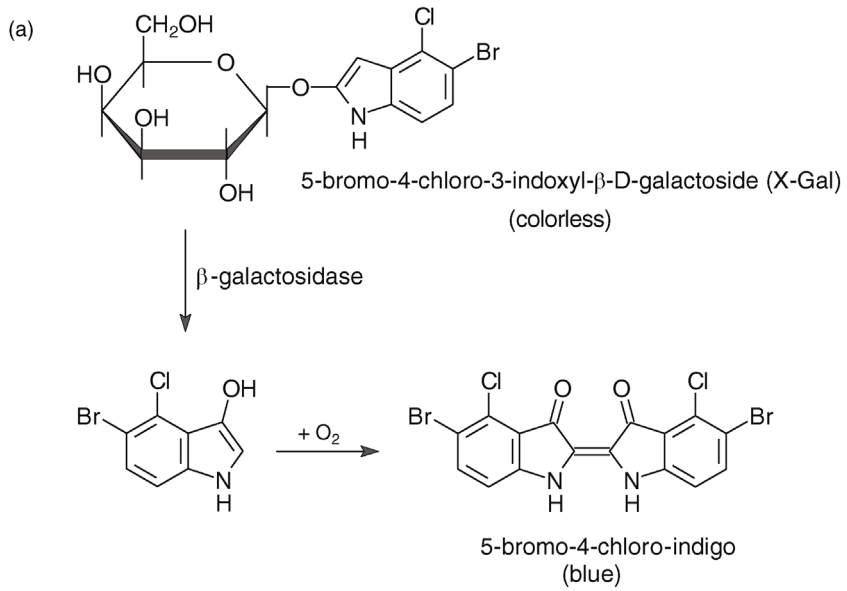


Figure 16.3 (a) The enzyme β-galactosidase (encoded by the lacZ gene) catalyzes the conversion of X-Gal (colorless) into 5,5′-dibromo-4,4′-dichloro-indigo (blue). (b) Histidine biosynthesis in yeast. A common reporter gene is HIS3 which is deleted in Y2H strains. Modified after pathway from the Yeast Genome Database (SGD). Four digit-numbers indicate Enzyme Commission (EC) numbers. AICAR = 5-Aminoimidazole-4-carboxamide ribonucleotide.

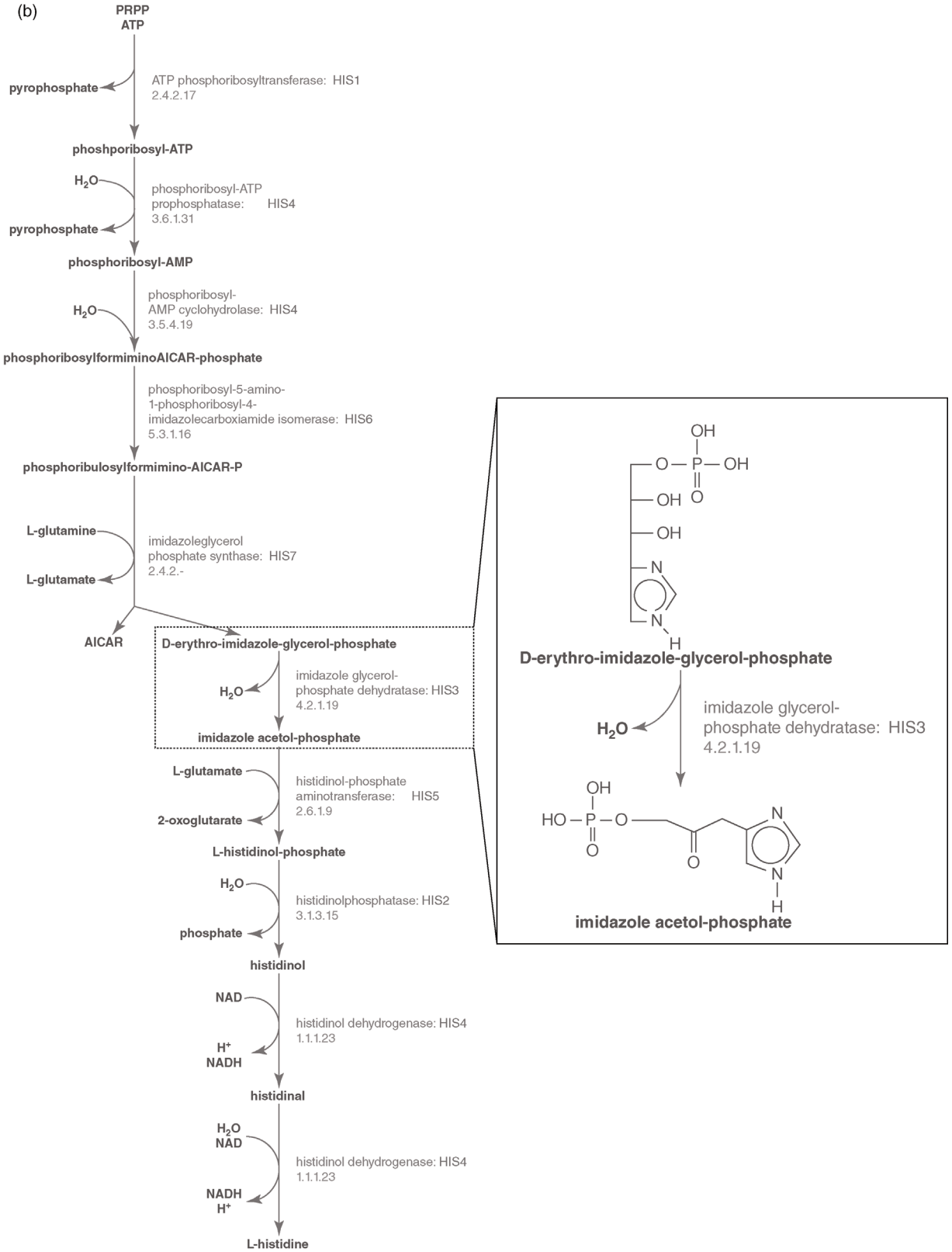


Figure 16.3 (*Continued*)

- Elements that drive propagation in yeast, namely, the 2μ origin of replication, which ensures replication of the plasmids when cells divide. The pBR322 origin of replication is active in *Escherichia coli* so that the plasmid can be propagated in bacteria, too.
- Several "marker" genes, including Trp1 and Leu2 that can be selected for in yeast. KanR and AmpR encode for resistance genes for kanamycin and ampicillin for selection in *E. coli*. That is, these plasmids are only maintained if the yeast cells are grown without tryptophan or leucine, or if the *E. coli* cells are grown in the presence of ampicillin or kanamycin antibiotics.

In addition, the bait protein needs to fulfil several criteria in order to be useful, namely:

- The fusion protein needs to be expressed, which is not always the case, for example, if it happens to be unstable or even toxic to the cell. Expression can be detected by a successful Y2H experiment or by other means, such as Western blotting.
- The fusion protein needs to be localized in the yeast cell nucleus. The DBD and AD fusions therefore also contain a nuclear localization signal (NLS), here from the simian virus 40 (SV40) genome.
- The fusion protein should not activate transcription in the absence of a prey protein. This actually happens more often than expected, in 10–20% of all cases.

The latter problem, "autoactivation," is fairly common, even if the bait is not a transcription factor. For instance, certain protein sequences such as those containing many acidic amino acids or glutamine-rich proteins can autoactivate. There are several ways to address this problem. If the reporter gene for the Y2H assay is the His3 gene, weak autoactivation leads to low level expression of the HIS3 protein, which can be suppressed by an inhibitor, 3-aminotriazole (3-AT) (Figure 16.4). This has the advantage that different levels of autoactivation can be suppressed by appropriate 3-AT concentrations.

If autoactivation is very strong, there is little that can be done. As a last resort, the autoactivating sequences may be identified and removed, that is, the bait protein can be split into segments that are screened individually. This is easier if functional segments are known, such as domains or transmembrane helices. On the plus side, segmentation can help to identify functional domains of a protein.

16.1.4 Which Bait Proteins can be used in a Y2H Screen?

Most proteins can be tested by two-hybrid systems although there are a few constraints. For instance, PPIs may depend on modifications such as phosphorylation, glycosylation, isoprenylation, and so on. If the yeast cell does not express the modifying enzyme, an interaction may not take place. However, expression of modifying enzymes may solve the problem, although this has been tried in only a few cases (e.g., kinases or histone acetylases).

Membrane proteins may not work either as they may not properly localize to the nucleus. Again, there is a ray of hope if the fusion protein is localized to the nuclear membrane. In fact, several transmembrane proteins have been successfully screened using the Y2H. In fact, specific Y2H systems have been designed for the use of membrane proteins (see below).

Similarly, extracellular proteins do not seem to work well in Y2H screens. However, the efficiency is difficult to evaluate as negative data is rarely published. In any case, extracellular proteins may mislocalize in the ER or Golgi before reaching the nucleus, but also may get modified by glycosylation or other modifications, diminishing their suitability to this assay.

16.1.5 AD Fusion Proteins and cDNA Libraries

Once the bait is made, the prey needs to be constructed from a defined protein or a library of clones. The prey is expressed from a separate plasmid and therefore can be selected independently. The prey vectors in Figure 16.2 have all the necessary elements for this, including different selective markers. In the case of pGADT7g the Leu2 gene is required for a cell to grow on leucine-free media (so the cells can make their own leucine). Note that the promoter and terminator for that gene are not shown in Figure 16.2.

Western-Blotting, Section 5.3, 11.11.3, 25.4.1

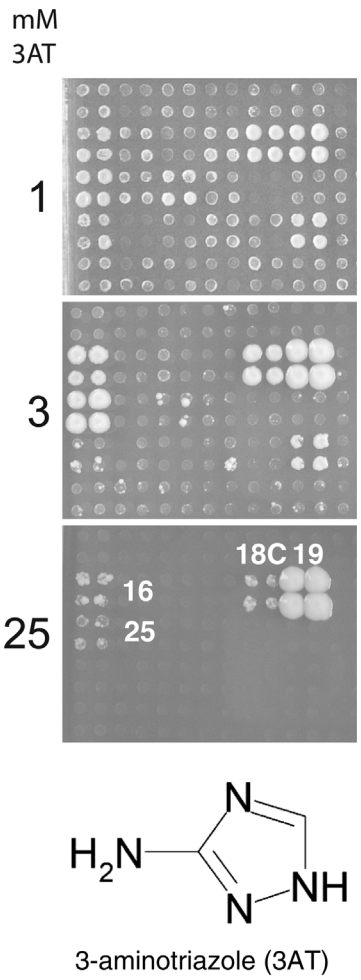


Figure 16.4 Inhibition of the HIS3 reporter protein by 3-aminotriazole (3-AT). Each photograph shows yeast colonies in groups of four, corresponding to a single interaction each. That is, each photograph shows an array of $5 \times 6 = 30$ protein pairs being tested. The figure shows interactions among the subunits of ribonucleotide reductase encoded by chickenpox virus (also known as the varicella zoster virus, VZV). The numbers in the third panel indicate the ORF numbers within the VZV genome. At 1 mM 3-AT a significant background signal is visible. At 3 mM the background is noticeably reduced and at 25 mM it is essentially gone. Note, however, that high concentrations of 3-AT may also eliminate weak interactions (like the one at the bottom right-hand side). In addition, 3-AT can be used to assess the strength of an interaction, even though Y2H assays are not truly quantitative. Source: adapted from Stellberger, T. *et al.* (2010) *Proteome Sci.* 8, 8.

As mentioned earlier, the cloning site is a Gateway cassette that can be replaced by defined inserts or whole libraries that have compatible recombination sequences. Note that pGADT7g and pGBKT7g also contain HA and myc tags, respectively, epitope tags that are recognized by specific antibodies for the detection of tagged proteins.

If a random cDNA library is inserted into the Gateway region (or by regular cloning) a continuous reading frame from Gal4 to insert is not guaranteed – only a third (or a sixth if the cDNA is not directional) of all clones are in the correct reading frame. If the library is comprehensive, this should not be a problem as most genes are represented in the library.

cDNA libraries or defined clone sets are now available for most model organisms (including humans) and their tissues. Defined clone libraries that contain all open reading frames (or genes) at the same ratio avoid the disadvantage of cDNA libraries which contain each gene only at the rate at which they are expressed, that is, some genes are highly overrepresented and others are not represented at all. In other words, such clone sets are *normalized*.

Gateway-Cloning, Section 16.1.3
GST, Section 16.3
GFP, Section 7.3.4, 8.4

The Gateway system, which is a commercial system, has the added benefit that numerous other vectors are available that enable the expression of a protein for many other purposes. For instance, once an ORF is in a Gateway entry plasmid it can be easily moved to a plasmid for expression as GFP (green fluorescent protein) or GST fusion that can be used to detect them microscopically or for their purification, respectively.

16.1.6 Carrying out a Y2H Screen

Prerequisites To carry out a Y2H screen, you need the bait plasmid(s), the prey clones or library, the yeast strains with reporter genes, and the media to grow yeast. The reporter genes may be integrated into the yeast genome or provided as plasmids (Figure 16.5). Note that the

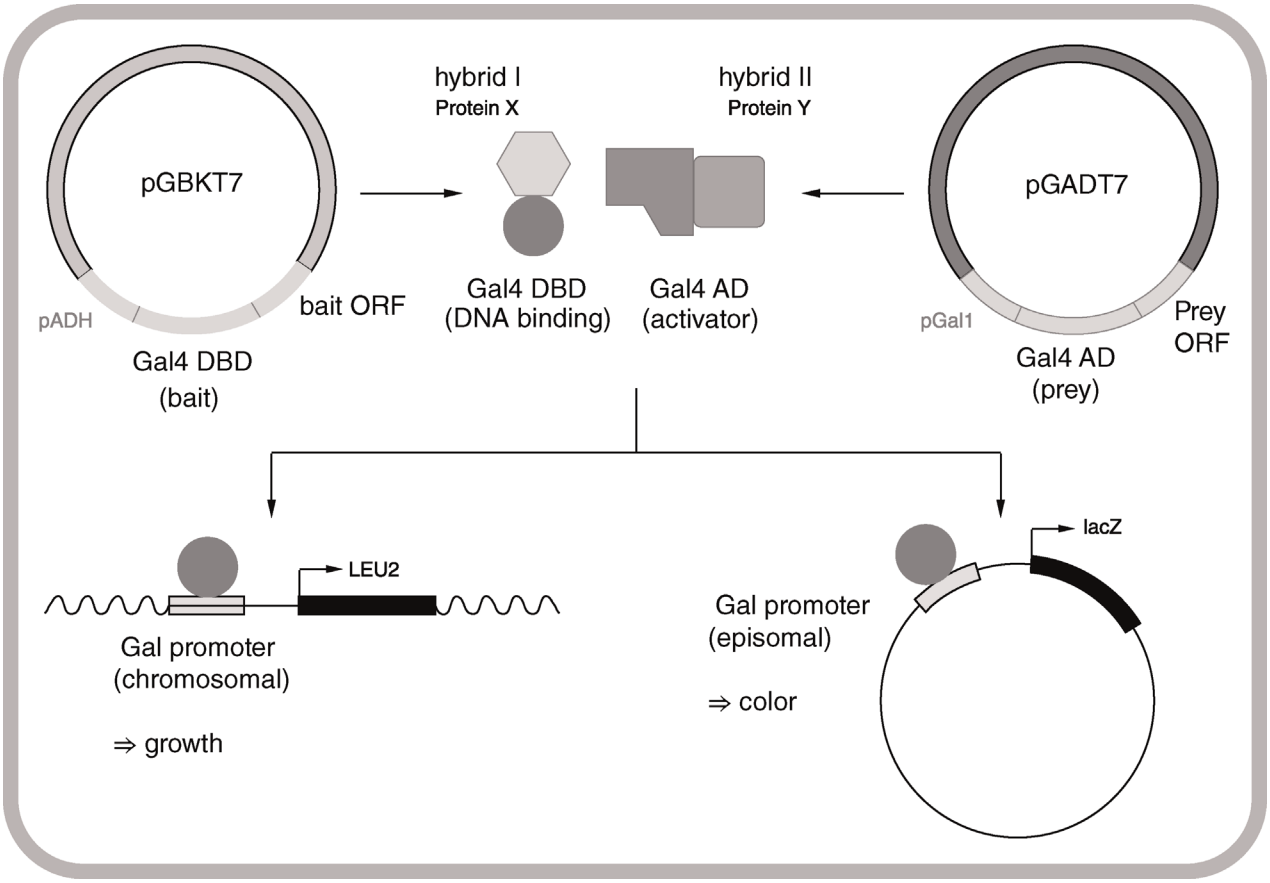


Figure 16.5 Protein–protein and protein–DNA interactions in a *Y2H* assay. Note that the reporter genes (here: LEU2 and lacZ) may be either on a chromosome or on a plasmid (i.e., "episomal").

yeast strain needs at least three reporter genes that complement their corresponding deletions, for example, of the Leu2, Trp1, and His3 gene, so that the Y2H plasmids can be selected, given that they provide these genes. The most commonly used marker genes are:

- Ura3 = orotidine-5′-phosphate decarboxylase (uridine synthesis),
- His3 = imidazole-glycerol-phosphate-(IGP-) dehydratase (histidine synthesis),
- Trp1 = phosphoribosyl-anthranilate isomerase (tryptophan synthesis),
- Leu2 = 3-isopropylmalate dehydrogenase (leucine synthesis).

The yeast strains AH109 and Y187 are often used for such screens, given that they have the required genotypes:

- AH109: MATa, trp1-901, leu2-3,-112, ura3-52, his3-200, gal4Δ, gal80Δ, LYS2::GAL1UAS-GAL1TATA-HIS3, GAL2UAS-GAL2TATA-ADE2, URA3::MEL1UAS-MEL1 TATA-lacZ
- Y187: MATα, ura3-52, his3-200, ade2-101, trp1-901, leu2-3,-112, gal4Δ, met$^-$, gal80Δ, URA3::GAL1UAS -GAL1TATA -lacZ

In this notation, (Δ) denote deletions, lower-case names are mutations (e.g. trp1-901), and double colons indicate insertions (e.g. URA3::GAL1UAS is a URA3 gene with an inserted Gal1 UAS etc.)

The two strains are auxotrophic for leucine, tryptophan, histidine, as well as uracil. Leu2 and Trp1 are only used to select for yeast that carry the two plasmids. His3 is used for the actual selection of cells that express interacting proteins. Ura3 and lacZ can be used as additional markers. In fact, different Y2H systems use different markers on different plasmids, hence you should keep an eye on the precise features of both plasmids and yeast strains.

If one strain has the bait plasmid, the prey plasmid(s) can be transformed into it in a separate reaction. However, more often the bait is transformed into Y187 and prey is transformed into AH109 independently, and then the two strains are simply mated to form diploid cells. *Mating* is usually much more efficient than transformation and takes place between the two "sexes" of yeast, called a und α, indicated by MATa and MATα in the genotype above.

The Interaction Screening The first and easiest step in a Y2H screen is to get the bait plasmid into the bait strain. The more difficult part is to transform the prey library or individual prey clones into the prey strain. Transformation must be highly efficient to obtain a good library. Many libraries are now commercially available and they are usually pre-transformed into yeast. If not, as a rule of thumb at least a million independent clones should be obtained in order to have rare cDNAs represented. For random libraries, having more clones is better because only a fraction of them are in the correct reading frame and many will contain only fragments of coding sequences; that is, a microgram of DNA should yield 10^5–10^6 transformants. This turned out to be difficult to achieve, given that yeast is much less "competent" than bacteria such as *E. coli,* which routinely produces 10^8–10^9 transformants per microgram of plasmid DNA. However, improved protocols for the LiAc–PEG transformation (lithium acetate–poly(ethylene glycol)) solved this problem to a large extent so that 10^5 transformants per microgram DNA are achieved in many laboratories today. Once a library has been obtained it can be frozen at –70 °C and samples from it reused multiple times.

As an alternative to cDNA libraries complete clone sets have been made for many species, that is, all ORFs have been cloned into standardized vectors so that they can be recloned into Y2H plasmids either as subsets or as a whole. Such clone collections are often called ORFeomes.

The Screening Historically, a cDNA library was often transformed into the bait-containing strain. This has been replaced by mating more recently. Here, a library of preys is mated to a bait strain by simply mixing the two cell types either in liquid or on an agar plate. While in close contact, the haploid yeast cells form diploid cells that contain both bait and prey plasmids. This step is more efficient than transforming cells but still only a small fraction of all cells mate. To get rid of the haploid cells, the mix is transferred to selective media that allow only those cells to grow that contain both bait and prey plasmids. In our case, this is achieved by growing the cells on media lacking leucine and tryptophan, as both strains have

mutations in Leu2 and Trp1, so they can grow only when these two genes are provided by the bait and prey plasmids.

Once the diploid cells have been selected this way, they can be transferred to the actual Y2H selection, in our case media that also lack histidine. Remember that a successful interaction of bait and prey proteins activates transcription of the *His3 reporter gene*, so only cells with interacting proteins can grow on media that lack histidine. When the mated cells are grown on such media ("-Leu-Trp-His" or simply "-LTH") only a few cells should grow. However, since there are often autoactivators present in libraries, sometimes many hundreds of colonies are found. While usually only a single defined bait is screened, in a random library screen the identity of the preys is unknown. Thus, colonies are picked and the prey identified by PCR amplification and DNA sequencing. If a defined set of preys is tested in an array format, the position in the array usually predetermines the identity of the preys and sequencing is not necessary. However, in this case the clones have to be identified first.

False Positives The Y2H system has often been criticized for the large number of false positives that it produces. False positives in a Y2H screen are those protein pairs that behave as if they are interacting when in fact they are not, for example, when a protein pair is identified in a yeast colony growing on selective media without the two interacting proteins actually inducing the reporter gene. The reason for such false positives is poorly understood but there are a couple of plausible mechanisms for how the reporter gene can be activated without an interaction. The most important one is autoactivation of a bait protein, hence autoactivators need to be identified before a screen, ideally by mating a bait strain with a prey strain that contains only an "empty" prey vector (i.e., the plasmid without prey ORF inserted). Sometimes preys can autoactivate, although the mechanism is less clear, for example, if a prey protein can activate histidine biosynthesis indirectly. Since it is virtually impossible to avoid false positives altogether in large screens, they have to be identified after the screen either experimentally or by computational analysis (see below).

Some false positives may be avoided by using additional reporters. For instance, you may replate positive colonies from an initial screen on X-Gal-containing media, using the lacZ reporter in strains Y187 and AH109. The cells then have to be lysed and β-galactosidase activity detected (see Figure 16.3). Some yeast strains contain additional reporters such as Ade2 or Mel1. Ade2 encodes phosphoribosylaminoimidazole carboxylase and catalyzes a step in the "*de novo*" purine nucleotide biosynthetic pathway; these cells can be selected in adenine-free media; in addition, mutations in ADE2 lead to the accumulation of purine precursors in the vacuole, which causes the colony to be red in color. Mel1 is a secreted α-galactosidase and is required for catabolic conversion of melibiose into glucose and galactose, hence the cells need to be grown on melibiose for selection. Note that these reporters can only be used if they have Gal4 binding sites (or any other binding site used by the DBD of the bait).

Another way to avoid FPs is to induce the hybrid proteins using inducers such as galactose. However, this only works if your bait and prey plasmids have galactose-inducible promoters. Note that the plasmids shown in Figure 16.2 do not have such promoters but rather constitutively active ADH promoters.

By growing these yeast cells on either glucose or galactose it can be tested whether it is really the expression of the bait or prey plasmid instead of a random mutation or some other effect that activates the reporter gene(s).

Re-testing and *Two-Hybrid* Arrays Instead of screening a random library, two-hybrid assays are increasingly used to test specific interactions that have been predicted by computational methods, for example, the homologous pair of a mouse interaction with human proteins. Similarly, if an interaction is known between proteins A and B, it may be worth testing all members if protein B is a protein family.

Such hypotheses can be easily tested with directed Y2H assays. In fact, instead of using a subset of proteins, all proteins of an organism can be tested that way. This can be done as a library screen where all clones are pooled, or in an array screen where all proteins are arrayed as shown in Figures 16.4 and 16.6. An array screen takes longer to set up than a library screen but the analysis of positives is much easier because the identity of each clone in an array is known, so that PCR amplification and sequencing is not necessary.

All preys are arrayed on an agar plate and then mated with a single (or sometimes multiple) bait strains. Arraying can be achieved manually by simply putting a drop of yeast cells on a

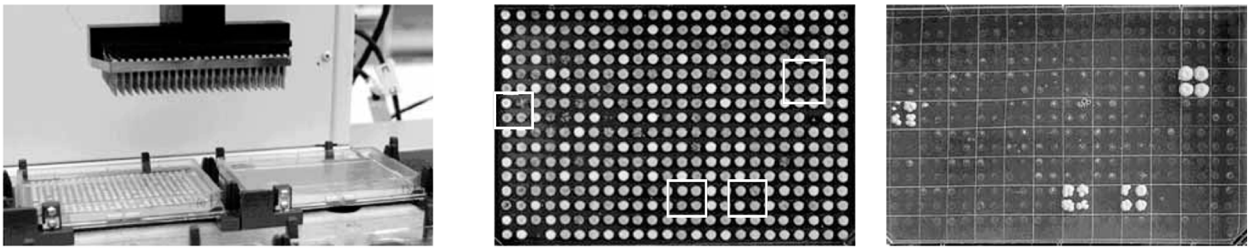


Figure 16.6 Array-screen: (a) Instead of a random library all 70 proteins of human chickenpox virus (VZV) were cloned into prey vectors and expressed in 70 × 4 yeast colonies (there are more than 280 colonies because some proteins were broken down into multiple fragments). Groups of four colonies contain identical proteins. Then, the same bait clone was stamped onto all prey colonies for mating. The next day the mated colonies were transferred to selective media. (b) After one week only those cells grow that express interacting proteins. The array format immediately reveals which proteins are interacting, how much background there is, whether the positives are reproducible, and how strong the interactions are (at least roughly). In this case a whole virus proteome was screened.

plate, or by "pinning" devices that pick up yeast from a 96- (or 384-)well plate and then put them on the agar plate. Robotic devices can automate the process to a large degree and thus process hundreds or even thousands of plates with little human intervention. Once the preys have been "pinned" down, the bait cells are put on top of them, so that both are in close contact. Usually, diploid cells form within a couple of hours, although mating is often carried out overnight to maximize efficiency.

As described for a library screen, the cells have to be transferred to selective media to remove un-mated cells and to verify successful mating. Finally, another transfer is usually made to put the diploid cells onto Y2H selective plates on which only those colonies grow that express interacting proteins (Figure 16.6).

Besides immediate identification of interacting protein pairs the array strategy is also more efficient in the identification of false positives because all pairs can be directly compared to each other (Figure 16.4). In this way the strength of interactions can be roughly assessed because "strong" interactions lead to higher reporter gene activity and thus more growth (i.e., larger colonies).

The colonies in Figures 16.4 and 16.6 are plated as quadruplicates, adding another level of control, making sure that an interaction is reproducible. In Figure 16.4 (middle panel) several single colonies indicate false positives as they are not reproducible in their full set of four.

Limits of the Two-Hybrid System A remarkable result from comprehensive Y2H screens was that they detected only 20% or 30% of all expected interactions, especially with proteins that had been extensively studied and were thus well understood. It turned out that this limitation actually applies to pretty much all assays, whether Y2H, *in vitro,* or *in vivo* assays (see below). When five different assays were compared in a study, they detected 20–36% of all interactions, and despite the sample was limited it was obvious that none of the methods was clearly superior to any other.

There are several reasons why PPI detection methods like the Y2H have such high false negative rates. Note that a success rate of 25% means that 75% of all true interactions are missed! First, the Y2H uses fusion proteins. If two proteins interact with surface areas that are anywhere near their N-termini, an N-terminally fused DBD or AD may block that interaction. This can be easily overcome by fusing the DBD and AD to the C-terminus and doing so indeed showed that many more interactions are detected (Figure 16.7).

Surprisingly, several other features of Y2H vectors seem to affect its sensitivity as well, such as the copy number of the plasmid, the promoter of the fusion proteins and thus their expression level, the linker between DBD and fused protein, and so on. When all of these factors are varied, up to 80% of all interactions may be detected. However, this requires the use of many different Y2H vector combinations, which will make many screens unacceptably complicated (Figure 16.7).

The Bacterial Two-Hybrid System Among the limitations of the Y2H system is the fact that it is carried out in the yeast nucleus. To detect PPIs both in membranes as well as outside of a nucleus a bacterial counterpart has been developed. The bacterial adenylate cyclase-based

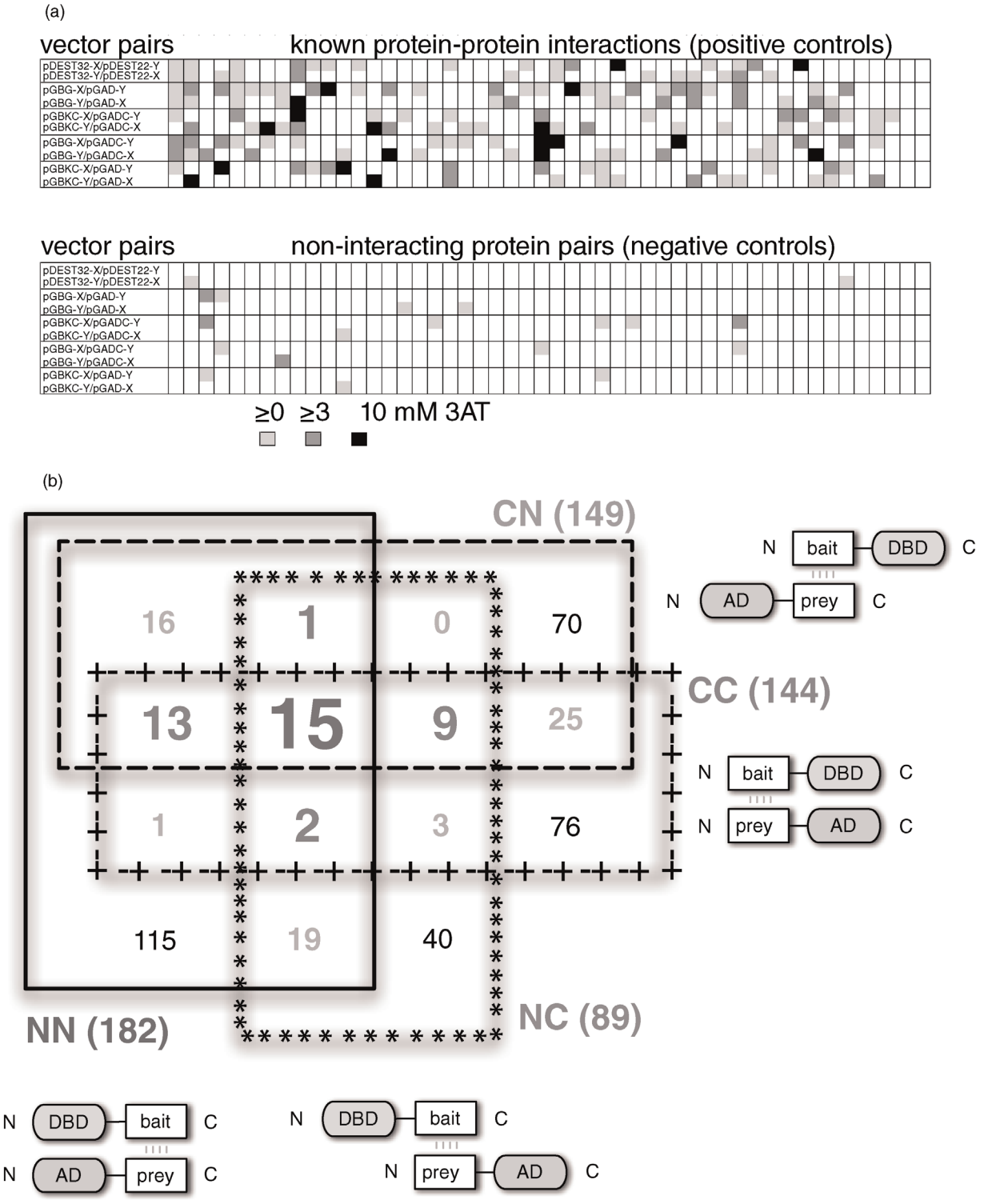


Figure 16.7 The limits of the Y2H system. (a) A single bait-prey vector pair can typically detect only 20–30% of all interactions of a certain bait protein. This limitation is caused by the sterical constraints of fusion proteins and also other features of the bait–prey pair. In this figure five bait–prey vector pairs were used to test the interactions of 50 well-known interactions (previously identified by a multitude of different methods, each column corresponds to one interaction). Each protein pair was cloned into both bait and prey vector so that the five different vector pairs yielded ten different bait–prey combinations (one per row). The interactions detected by these vector permutations not only differed in what was detected but also in their sensitivity to 3AT: many interactions were rather weak (light colors) or only suppressed by high 3AT concentrations (darker colors). Note that few interactions were detected by all vector pairs and some interactions were not detected at all (i.e., by none of the vector pairs). The lower panel shows non-interacting protein pairs (as far as known), which served as negative controls. They show how many false positives should be expected for each vector pair. (b) Similar to (a), four vector pairs were compared in a study of chickenpox virus proteins. Note that different Y2H vectors produced both N- and C-terminal fusion proteins ("hybrids") that were combined in Y2H tests. Each combination produced a different, non-overlapping set of interactions. Only 15 interactions were found with all four combinations. Source: (a) after Chen, Y.C. *et al.* (2010) *Nat. Methods*, 7, 667; (b) from Stellberger, T. *et al.* (2010) *Proteome Sci.* 8, 8.

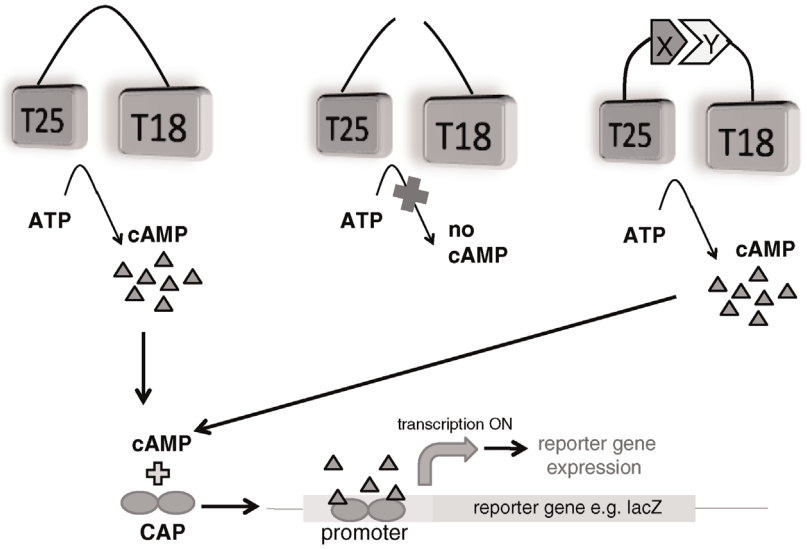


Figure 16.8 The bacterial two-hybrid system (BACTH or B2H). The BACTH is based on the re-association of the enzyme adenylate cyclase which has been split into two halves called T18 and T25. The two fragments cannot associate on their own but when they are fused to two interacting proteins (X and Y) a functional adenylate cyclase (AC) is reconstituted. AC converts ATP into cAMP, which in turn activates the catabolite activator protein (CAP) to become an active transcription factor. As in the Y2H a reporter gene is used to detect the interaction of X and Y.

two-hybrid system (BACTH) is based on the functional reconstitution of an adenylate cyclase activity in a cya^{-} *E. coli* strain. The catalytic domain of bacterial adenylate cyclase has two subdomains, a 25 kDa fragment (T25) having a catalytic site and a 18k Da fragment (T18) that has a binding site for calmodulin. Both domains are non-functional when they are physically separated (Figure 16.8). The co-expression of these fragments in the presence of calmodulin can reconstitute fully functional adenylate cyclase, resulting in cAMP synthesis. Thus, the BACTH is based on the functional complementation of these two subdomains of adenylate cyclase. Two test proteins X and Y can be fused to these fragments T25 and T18. The interaction between these proteins can result in functional complementation between the T25 and T18 fragments and eventually cAMP synthesis. The cAMP produced by the reconstituted adenylate cyclase forms a complex with catabolite activator protein (CAP), turning on the expression of several genes, including the lac and mal operons. The activation of lac or mal operons makes the bacteria competent to utilize lactose or maltose as carbon sources, which can be easily identified by different phenotypes on specific media plates (Figure 16.8).

A BACTH screen requires the construction of fusion test proteins with T25 and T18 fragments, similar to the bait or prey clones in Y2H plasmids. Then, the test proteins are co-expressed in adenylate cyclase lacking bacteria (*E. coli* cya^{-}) to screen for true interactions. The co-transformation of bacteria (e.g., strains BTH101 or DHM1) by recombinant plasmids can be carried out by standard transformation methods. The co-transformants can be selected and screened for interactions on indicator plates (Figure 16.8). Conveniently, the detected protein–protein interactions can be quantified: the cAMP/CAP complex formation positively regulates the expression of β-galactosidase and thus the cAMP levels in bacterial cells can be measured by a β-galactosidase assay. Furthermore, additional methods such as immunoprecipitation, LUMIER assay, copurification, and so on should be used to further characterize and confirm the detected interactions (Table 16.2).

16.1.7 Other Modifications and Extensions of the *Two-Hybrid*-Technology

The principle behind the Y2H system represents a much more fundamental idea, namely, that proteins can be split and their reassociation used to detect the interaction of a second pair of proteins. This principle is generally called "protein fragment complementation." While in Y2H a transcription factor has been split and reassociated by two interacting proteins, many other proteins can be split in a similar way (Table 16.1).

For instance, one of the first split proteins after Gal4 was the yeast ubiquitin (Ub) protein (Figure 16.9a). Such variations of the Y2H assay often have advantages compared to the

Table 16.1 Proteins used in protein fragment complementation assays (PCAs).

Abbreviation	Protein
DHFR	Dihydrofolate reductase
Amp	β-Lactamase
Gal4	Transcription factor Gal4 (yeast)
TEV	Tobacco etch virus protease
Luc	Luciferase
Ub	Ubiquitin
GFP	Green fluorescent protein
LacZ	β-Galactosidase
IFP1.4	Infrared fluorescent protein (*Deinococcus radiodurans*)
FAK	Focal adhesion kinase

transcription based system. The split-Ub system, for example, uses interactions that take place in the cytoplasm. Hence, proteins do not need to be localized to the nucleus. Since split-Ub is not based on transcription, there are no false positives that result from artifactual transcriptional activation through acidic proteins and the like. Hence, split-Ub provides an alternative system to analyze true transcription factors and their interactions.

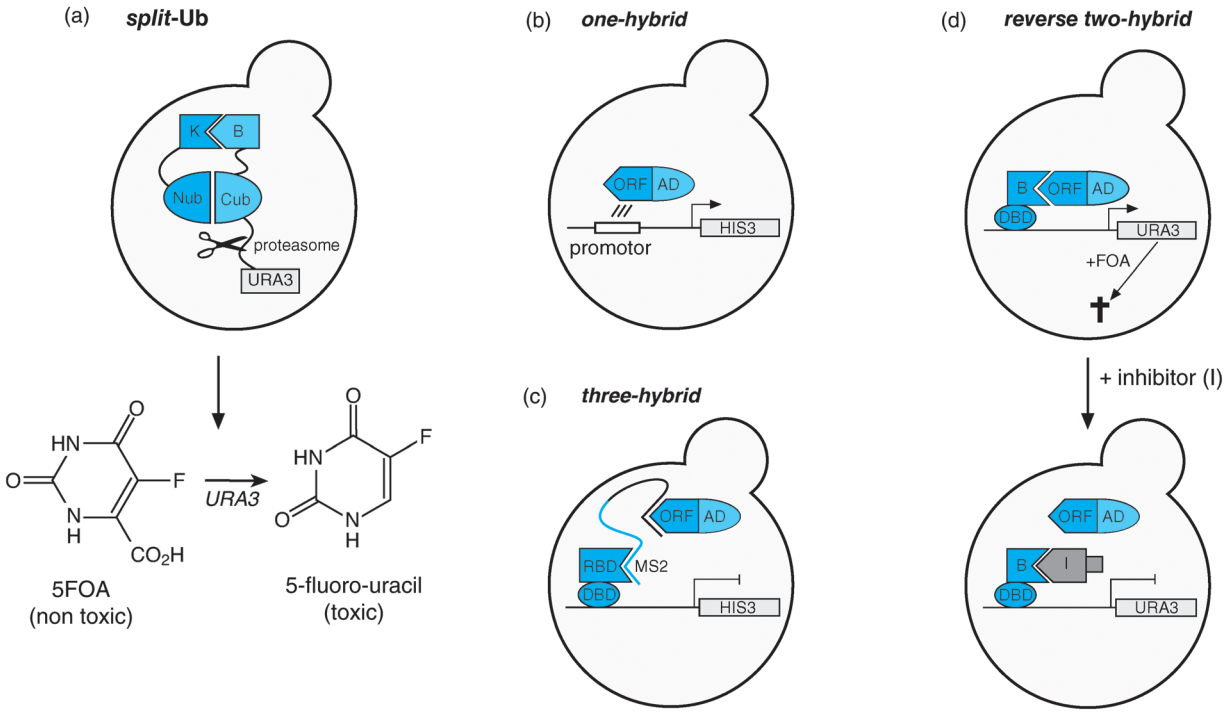


Figure 16.9 Modifications of the two-hybrid system. (a) The *split*-ubiquitin system uses reassociated halves of ubiquitin instead of a split transcription factor. If bait and prey interact, the N- and C-terminal fragments of ubiquitin (Nub and Cub) reassociate and thus can be recognized by the proteasome, which leads to degradation of the fused Ura3. Without an interaction Ura3 will not be degraded. In the presence of 5-FOA (5-fluoro-orotate) only cells *without* Ura3 (i.e., without a PPI) survive because Ura3 converts 5-FOA into the lethal metabolite fluoro-uracil. (b) The one-hybrid system uses a single fusion of an activation domain to any ORF, which may or may not bind to the promoter element of a reporter gene (here: His3). DNA-binding domains and promoter elements can be identified with such a system. (c) The three-hybrid system is used to find RNA-binding proteins. A DNA-binding domain (e.g., from Gal4 or LexA) is fused to an RNA-binding domain that recognizes a known RNA sequence (e.g., the MS2 RNA of phage MS2). The RNA must be a hybrid with another, uncharacterized RNA, binding to the second fusion with another RNA-binding protein of unknown specificity. (d) The reverse *two-hybrid* system uses 5-FOA as selection mechanism as in the split-Ub system: only interactions that are *blocked* prevent the synthesis of Ura3 and thus the lethal metabolite 5-fluoro-uracil. The system can be used to find compounds that *block* an interaction.

On the other hand, transcription factors can also be studied using a one-hybrid system. If a transcriptional activation domain is fused to a library of uncharacterized proteins a reporter gene can be cloned downstream of a suspected regulatory sequence and the proteins identified that bind to these sequences (Figure 16.9b).

Another variation can be used to detect RNA–protein interactions (Figure 16.9c). The DBD domain is fused to a known RNA-binding domain, for example, the domain that binds to the MS2 RNA. The Ms2 RNA is then fused to an uncharacterized RNA (expressed from a fused DNA fragment). The AD is fused to a library of ORFs which is screened for proteins that bind to the second RNA fragment. In this way RNA-binding proteins can be found. Technically this is a three-hybrid system, consisting of two protein fusions and an RNA fusion.

Finally, the two-hybrid system can also be used to find small molecules or drugs that block (or affect) protein–protein interactions. For instance, yeast can be modified in such a way that it only grows if a PPI is blocked. This is achieved by having a reporter gene that kills the cell when expressed. Ura3 has this property, but only if the cell is grown in the presence of FOA, which is converted into 5-fluoro-uracil, which is toxic to yeast (Figure 16.9d). Hence the system is called the "reverse two-hybrid system" because it detects the *absence* of interactions.

16.1.8 Biochemical and Functional Analysis of Interactions

Now that you have found some interactions the difficult part starts: Which one is really relevant and why? A two-hybrid interaction does not tell us whether it really takes place in a cell and what its biological role is. For instance, we may find an interaction between two proteins that never meet each other in a cell, for example, if they are expressed in different tissues or in different compartments of the cell. On the other hand, such interactions may be particularly interesting! A spectacular example is the interaction between the ER transmembrane protein SREBP (sterol response element-binding protein) and SCAP, which mediates the transport of SREBP to the Golgi apparatus. At low cholesterol concentrations both proteins are localized in the Golgi apparatus and the cytoplasmic domain of SREBP is cleaved off. The soluble cleavage product is translocated into the nucleus and acts there as a transcription factor, interacting with other components of the transcriptional machinery. In this case, a transmembrane protein of the ER and Golgi interacts in a physiologically relevant way with nuclear proteins! Interestingly, at high cholesterol concentrations both SREBP and SCAP proteins remain in the ER and no cleavage occurs. There are a number of other processes in which proteins of different compartments interact. There are several strategies with which to address these questions.

Independent verification In typical Y2H screens multiple interactions are found. Which one is relevant? A first attempt to separate the wheat from the chaff is to confirm an interaction with an independent assay. For instance, the two proteins may be copurified, using a GST-pulldown experiment (Section 16.3). However, this requires that at least one protein is re-cloned into a different vector that expresses a GST fusion protein, for example, in *E. coli*.

Sequence Data Analysis, Chapter 33

Analysis of Promoter Strength and Nascent RNA Synthesis, Chapter 34

A whole battery of assays are now available for such confirmation experiments. A few are listed in Table 16.2. Others are described in subsequent sections.

Table 16.2 Assays for the confirmation of PPIs.

Abbreviation	Full term	Reference
LUMIER	Luminescence-based mammalian interactome mapping	Barrios-Rodiles, M, Brown, K.R., Ozdamar, B., Bose, R., Liu, Z., Donovan, R.S., *et al.* (2005) *Science*, 307 (5715), 1621–1625
LumPIS	Luminescence-based maltose-binding protein pull-down interaction screening system	Vizoso Pinto, M.G., Villegas, J.M., Peter, J., Haase, R., Haas, J., and Lotz, A.S. *et al.* (2009) *Proteomics*, 9 (23), 5303–5308.
NAPPA	Nucleic acid programmable protein array	Ramachandran, N. *et al.* (2008) *Nat. Methods*, 5, 535–538.
PCAs	Protein fragment complementation assays	Remy, I, Campbell-Valois, F.X., and Michnick, SW. (2007) *Nat. Protocols*, 2 (9), 2120–2125

Computational Analysis of Interaction Data If a prey protein has a well-known function the biological meaning of an interaction may become obvious immediately. However, often prey proteins have no obvious function, so that their sequence or other data must be investigated for functional cues. There are many bioinformatics tools to analyze the properties of interacting proteins. For instance, hydrophobic sequences indicate transmembrane or secreted proteins. In bacteria, the location of a gene in an operon suggests a related function with those joined genes. Numerous databases can be searched for expression patterns of proteins, in particular organelles, cells, tissues, or species, and sometimes detailed quantitative information is available about time courses and levels of expression, both at the level of mRNA or proteins. In fact, large-scale projects have identified protein complexes in several model organisms, including *E. coli*, *Mycoplasma*, yeast, and human.

Biological Relevance The key question though is whether an interaction is of biological relevance. Although in most cases bioinformatics analysis will be able to make functional predictions, sometimes interacting proteins are completely uncharacterized and there are not even sequence homologies. However, some expression databases should have limited data and that can be used to design an experimental analysis. As an example, if the expression level is low or non-detectable, it is likely that the gene is only expressed under special circumstances. For instance, proteins may only be detected in cells that are starved of certain nutrients or under certain stresses, for example, high temperature, salt, or pH, or in the presence of toxins, antibiotics, or other compounds. It is also possible that a gene is involved in the synthesis of a compound that is only required under certain conditions, for example, when the cell starves and before it forms spores.

A powerful way to test such hypotheses is to make a mutation or, even better, just obtain a mutation form one of the numerous large-scale mutagenesis projects. Many, if not all, genes have been mutated in most model organisms, ranging from several bacteria to yeast, *Arabidopsis,* and mouse. These mutants can be tested for phenotypes under specific conditions.

Proteins may also be overexpressed, causing either a phenotype or less obvious change. If an overexpressed (mutated) protein displaces a protein from its natural complex that may have a dominant-negative effect.

Protein Domains and Motifs Most proteins consist of multiple functional units. For instance, protein kinases contain a kinase domain but also a binding domain that couple them to their target proteins or certain locations within the cell. Even if there is only one domain, often there are shorter sequences or even individual amino acids such as binding sites or catalytic centers in an enzyme. Binding sites for other proteins can be mapped using the Y2H, for example, by using fragments of a protein or mutated proteins. Such studies can and should always be supported by computational tools, especially when the structures of the proteins in question are available.

Localization Interacting proteins should co-localize. This can be tested by immunochemistry, using antibodies that detect both proteins in a cell. Alternatively, co-immunoprecipitation can show that proteins associate *in vivo*. If antibodies are not available, epitope tags need to be attached to the two proteins for detection. A more dynamic analysis of PPIs is possible by FRET analysis (Section 16.7).

16.2 TAP-Tagging and Purification of Protein Complexes

Tilmann Bürckstümmer

Horizon Discovery Ltd, Campus Vienna Biocenter 3, 1030 Wien, Austria

Mass Spectrometry, Chapter 15

Tandem affinity purification (TAP) is a method used to purify protein complexes from cells or tissues and to determine their composition by *mass spectrometry* (Figure 16.10). The bait protein (following the terminology of the two-hybrid system) is fused to a TAP-tag and expressed in a cell line or even a transgenic organism. The protein complex is then purified using the two tags (hence

"tandem") followed by mass spectrometric analysis. The method is similar to co-immunoprecipitation and GST pulldown (described below) but has several *advantages:*

1. In contrast to co-IPs TAP is unbiased due to MS analysis: no a priori assumption is needed as to what will bind to a bait. In a co-IP only those interactors can be detected for which antibodies are available (although MS can be applied to co-IPs, too). However, gene expression in the target cells is limiting as only those interactors will be found that are expressed.
2. Compared to other affinity purification tags (His_6, FLAG, GST) TAP works well even when the bait protein is expressed at very low levels and in a complex mixture such as a mammalian cell lysate. This is achieved by the high affinity of the tag for its matrix (see below).
3. Compared to Y2H assays TAP can purify complexes in their entirety while Y2H detects only binary interactions. This is a significant advantage if multiple components are interacting in a cooperative way, that is, when more than two proteins are required to stably interact.
4. In contrast to Y2H, TAP can detect interactions that are dependent on post-translational modifications such as phosphorylation and ubiquitination. In yeast these modifications often do not take place (or at a lesser extent or specificity).

TAP does usually require a stable cell line that expresses the TAP-tagged bait. While several methods are available to make such lines we only describe retroviral transduction here because it is a good compromise between efficiency and safety. If you have cells that are easily transfected, you may also use plasmids for transfection (or lipofection) and subsequent selection using antibiotics. Primary cells which only divide slowly (if at all) cannot be transfected by retroviruses. Lentiviral transduction is usually the best way to introduce expression plasmids into such cells.

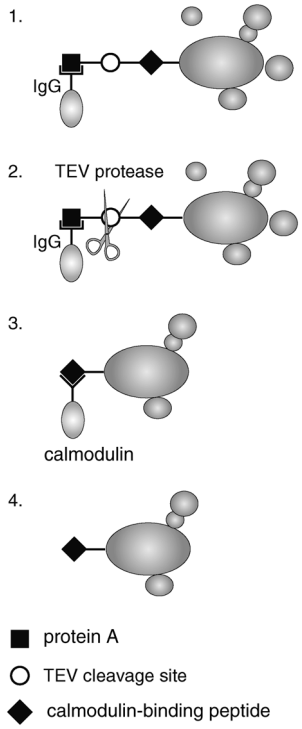


Figure 16.10 Tandem affinity purification (TAP).

Retroviral Transduction

Transfer of DNA into Mammalian Cells, Section 34.4.2

Recombinant retroviruses are a good way to transducer most cell lines. This requires a retroviral plasmid that encodes the TAP-tagged bait protein. The plasmid also requires retroviral long-terminal repeats (LTRs) for the packaging of retroviral mRNA into virus capsids. Ecotropic retroviruses infect only rodent cells while amphotropic retroviruses infect both rodent and human cells. The tropism of a retrovirus is determined by the glycoprotein on the virus surface which binds to target cell. For ecotropic viruses the Moloney murine leukemia virus envelope (MMLV Env) protein is used because it binds to the murine mCAT1 receptor. Amphotropic viruses are used with the G protein of vesicular stomatitis virus (VSV-G), which binds to mammalian phosphatidyl-serine. For safety reasons amphotropic virus are often avoided as they also infect human cells. To reduce the risks even further, viruses can be rendered replication deficient and thus will survive only one infection cycle.

To infect human cells with recombinant retroviruses, a retroviral packaging cell line is used (e.g., 293gp) to produce recombinant viruses. This cell line expresses the viral polyproteins Gag and Pol which are required in the virus structure. The packaging cell line is transfected with two plasmids, including one with the TAP-tagged transgene and one encoding the viral glycoprotein (VSV-G). Recombinant viruses are produced and accumulate in the supernatant within 48 h and they can be used to transfect the target cells. Infection usually only reaches a fraction of the cells, hence transfected cells may be selected either by an antibiotic resistance gene or GFP (encoded by the retroviral transfer plasmid). The result is a pool of stable cells that express a TAP-tagged bait protein. Sometimes it may be required to isolate a monoclonal cell population before the cells are used for TAP. Depending on the method, an experiment requires 2×10^7 to 10^8 cells. TAP experiments thus require a relatively high investment in cell culture and are time-consuming and costly.

TAP Purification The original TAP protocol was developed by Bertrand Séraphin's group at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany, in the late-1990s. They constructed the TAP-tag that consists of two immunoglobulin-binding domains of *Staphylococcus aureus* Protein A and the calmodulin-binding peptide (CBP). Both are separated

by a peptide that is recognized and cleaved by Tobacco etch virus (TEV) protease (Figure 16.10). Protein A binds immunoglobulines and *S. aureus* exploits this fact to avoid recognition by macrophages. CBP is a short motif that binds calmodulin in the presence of calcium ions (Ca^{2+}).

During the first step of a TAP purification the bait fusion protein is purified by binding to a matrix that is covered by antibodies (immunoglobulins, Igs). Protein A binds to Igs with high affinity (at nanomolar levels) and thus can also purify complexes of very low concentration. This is significant because other purification systems, such as GST or His_6, have a lower affinity for their affinity reagents than Ig and thus would be able to purify only those bait proteins that are present in fairly large amounts. However, such conditions are not compatible with the purification of intact and physiologically relevant protein complexes because their binding partners are only present at low levels and the bait protein may otherwise primarily bind highly abundant proteins such as heat shock proteins.

After enrichment of the protein complex on the immunoglobulin matrix the TEV protease is added to release the complex from the matrix. TEV protease is particularly suited for this process as the first elution needs to be carried out under mild conditions so that the complex is not disintegrated by harsh purification. In addition, TEV protease is highly specific and most proteomes, including the human proteome, have only few TEV recognition sequences, hence only a few human proteins are cleaved and destroyed by this protease.

In a second purification step the bait protein (and associated proteins) are enriched by means of a calmodulin matrix. Since binding requires Ca^{2+}, complexes can be easily eluted from that matrix by using a metal chelator such as EGTA. Alternatively, the sample can be simply boiled in SDS sample buffer.

In the meantime several TAP variants are available that have improved the efficiency of TAP purifications (Table 16.3). However, the principle is very similar to the original idea, so we will not discuss those alternatives in more detail.

TAP purifications were originally developed for yeast cells, which allows for efficient recombination and production of C-terminally tagged bait proteins *within* the yeast genome. Since then nearly all yeast genes have been TAP-tagged and these complexes purified. Similar large-scale projects have been carried out with *E. coli* and human cells although the latter

Table 16.3 Commonly used epitope tags for the affinity purification of proteins.[a)]

"Simple" tags	Affinity reagent
Protein A/Protein G	Immunoglobulin (IgG)
Glutathione S-transferase (GST)	Glutathione-Sepharose
Maltose-binding protein (MBP)	Amylose
Green fluorescent protein (GFP)	anti-GFP
Streptavidin-binding peptide (SBP)	Streptavidin
Calmodulin-binding peptide (CBP)	Calmodulin
HexaHis (His_6) tag (HHHHHH)	Nickel NTA
HA tag (YPYDVPDYA)	anti-HA
V5 tag (GKPIPNPLLGLDST)	anti-V5
Myc tag (EQKLISEEDL)	anti-Myc
FLAG tag (DYKDDDDKG)	anti-FLAG
Strep tag (WSHPQFEK)	StrepTactin
TAP tags (tandem-tags)	**Reference**
Protein A- TEV- CBP	Rigaut, G. *et al*, (1999) *Nat. Biotechnol.*, 17, 1030
Protein G- TEV- SBP	Burckstummer, T. *et al*. (2006) *Nat. Methods*, 3, 1013
LAP tag	Cheeseman, I.M. *et al*. (2005) *Sci. STKE*, (266), 1
Strep-HA	Glatter, T. *et al*. (2009) *Mol. Syst. Biol.*, 5, 237

a) The amino acid sequence is only provided for short tags. For the tandem tags literature references are provided that should be consulted for details. See also Section 16.3 for GST pulldowns.

required retroviral transfection as described above. The purification of human complexes turned out to be trickier than yeast or bacteria, not the least because the human proteome is far more complex. In addition to the retroviral transgene this approach may suffer from the presence of the endogenous bait protein, which competes for binding partners. Hence a slight over-expression is actually beneficial for successful TAP purifications.

Mass Spectrometric Analysis The product of a TAP purification, ideally a mixture of interacting proteins of low complexity, is typically digested with a protease, usually trypsin. The resulting peptides can be separated by liquid chromatography (Section 10.4) and then identified by mass spectrometry (Chapter 15). The sensitivity of MS analyses of peptides has been dramatically improved over recent years. However, it still suffers sometimes from low dynamic range, that is, the mass spectrometer selects the most abundant peptides for sequencing. Therefore, it is important that the purification contains as little contamination as possible. Otherwise highly abundant proteins such as heat shock proteins or ribosomal proteins may be sequenced instead.

This is where the two-step ("tandem") purification comes in as single-step purifications rarely reach sufficient purity. If necessary, the product of a TAP purification can be further separated by SDS PAGE (Section 11.3.9) or HPLC (Section 10.4) in order to identify underrepresented proteins, too.

Limitations of TAP Purification

1. TAP purifications are rather time-consuming and cost intensive, requiring access to sophisticated cell culture techniques (retroviral transfection) and mass spectrometry.
2. TAP purifications are usually able to detect robust interactions very reliably. However, transient and weak interactions are trickier to detect. Often, such interactions do not survive the purification process.
3. TAP purifications can isolate complexes from cells under nearly physiological conditions. However, the result is a list of interacting proteins without information as to which proteins interact directly with each other. Additional experiments such as co-IPs or yeast two-hybrid experiments are required to answer these questions.

16.3 Analyzing Interactions *In Vitro*: GST-*Pulldown*

Peter Uetz

Virginia Commonwealth University, Center for the Study of Biological Complexity, 1000 W Cary St, Richmond, VA 23284, USA

While affinity purification by TAP tagging can isolate complexes from cells it is often necessary to analyze selected interactions or complexes *in vitro*. In fact, *in vitro* analysis allows us to investigate interactions in isolation from other (unwanted) proteins. This is why our target proteins should be expressed in a heterologous system, for example, human proteins in *E. coli* or yeast just to avoid co-purifying proteins or modifications by enzymes.

A well-established method for *in vitro* interaction analysis uses fusion proteins of glutathione-S-transferase (GST) (Figure 16.11). GST fusion proteins can be easily purified from *E. coli* by running a cell extract over a matrix coated by glutathione. Usually tiny glutathione-coated beads made of Sepharose are used for that, so they can easily be used in a test tube. When these beads are incubated with an extract that contains GST fusion proteins, only those fusions and their associated proteins bind to this matrix. Unspecifically bound proteins can be removed by washing the beads in a salt-containing buffer such as PBS (phosphate-buffered saline). Subsequently the fusion proteins can be eluted by incubating the beads in a solution of glutathione which competes with the matrix bound glutathione for the fusion protein. If necessary, the glutathione can be removed by dialysis (see section 2.6).

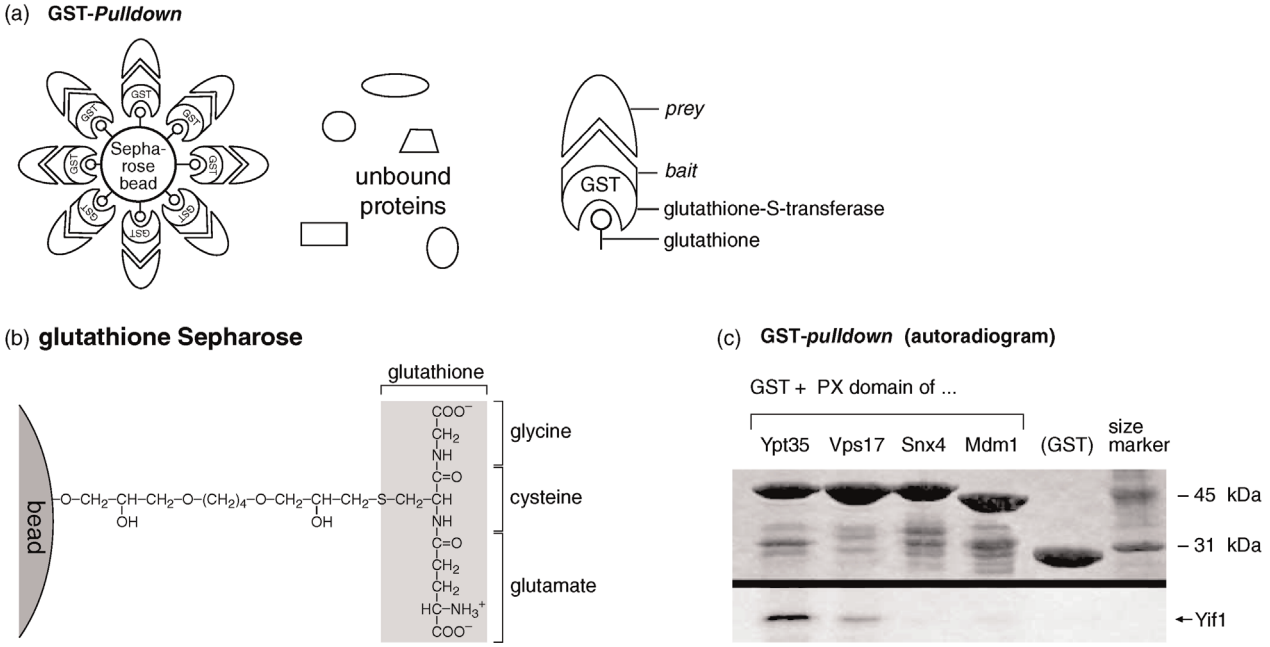


Figure 16.11 GST-pulldown: (a) the principle (see text for details). (b) Structure of glutathione Sepharose. Glutathione is a natural tripeptide that also occurs as an oxidized dimer when the SH groups of two molecules react. (c) Example: the co-called PX domains of four yeast proteins (Ypt35, Vps17, Snx4, Mdm1) were fused to GST and purified. The gel pictures (top) show those proteins stained by Coomassie Blue (all around 45 kDa in size). The beads with bound fusion proteins were then incubated with yeast protein Yif1 that had been synthesized and radioactively labeled with ^{35}S-methionine *in vitro*. After washing, bound protein was released from the beads by boiling in sample buffer, separated on a gel, and detected by exposing the dried gel to X-ray film (bottom). The experiment shows that Yif1 only binds to the PX domains of Ypt35 (and weakly to Vps17) but not those of Snx4 or Mdm1. Source: modified after Vollert, C.S. and Uetz, P. (2004) *Mol. Cell. Proteomics*, 3, 1053–1064.

Usually, it is easier to leave the fusion protein on the matrix without eluting it, so the beads can be directly incubated with a second protein (or mixture) to measure its binding to the fusion protein. The binding protein may be radioactively labeled by *in vitro* transcription and translation of a cloned gene. This can be easily achieved by commercial kits that contain all reagents, RNA polymerase, and even an extract with ribosomes and other factors. Only a PCR fragment or plasmid that encodes the protein of interest and radioactively labeled ^{35}S-methionine have to be added. As long as the DNA contains a promoter for the RNA polymerase (typically from phage T7) the transcription and translation happens *in vitro*. The only limitation is that the cloned protein needs to encode at least one methionine so that the radioactive label will be incorporated into the protein. However, as an alternative, an epitope-tagged protein can be used as well.

Western-Blotting, Section 5.3.3, 11.7

The *in vitro* translated protein is then mixed with the GST-protein (bound to beads) and incubated. Then the beads need to be washed and ideally only the GST-fusion and its bound proteins are retained on the beads. Most commonly, the beads will then be boiled with sample buffer and separated on a SDS-PAGE gel (Figure 16.11). If the sample contains enough protein it can be detected by Coomassie Blue or some other staining (Figure 16.11). Detection in a gel also reveals its size (provided a suitable marker is included). If the amounts of protein are too small the gel can be blotted onto a membrane where proteins can be detected by antibody staining (e.g., using enhanced chemiluminescence, ECL, Section 16.5), radioactive labelling, or other methods.

Polyacrylamide Gel Electrophoresis, Section 11.3.9

16.4 Co-immunoprecipitation

Immunoprecipitation, Section 5.3.2

Co-immunoprecipitation (co-IP) is very similar to GST-pulldowns (Figure 16.12). However, instead of glutathione Sepharose co-IPs use Sepharose beads that are coated by protein A from *Staphylococcus aureus,* which binds to the constant chains of IgG antibodies with very high affinity. Hence, protein A-coated Sepharose beads can be easily coated with antibodies. These beads can then be incubated with proteins that bind to the antibodies, for example, from cell or tissue extracts.

Mass Spectrometry, Chapter 15

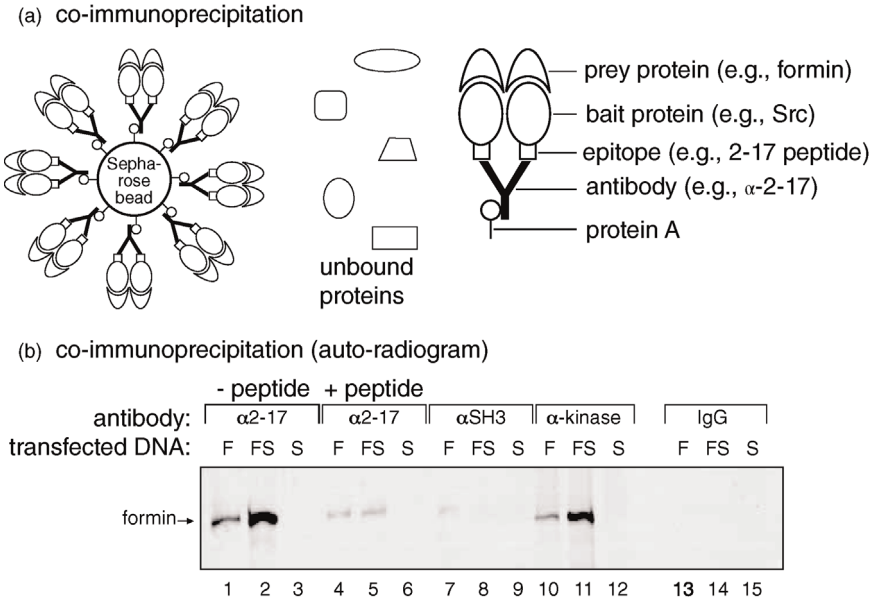


Figure 16.12 Co-immunoprecipitation: (a) principle (see text for details). (b) Example: a so-called formin protein was co-precipitated with the oncoprotein Src. In this experiment, four different anti-Src-antibodies were used: one that bound to amino acids 2–17 of Src (α2–17), one against the SH3 domain (αSH3), one against the kinase (α-kinase), and a control antibody without known specificity (IgG). The interacting proteins "formin" and "Src" were either expressed separately (F or S) or together (FS) in tissue culture. The cells were lysed and incubated with the antibodies coupled to protein A–Sepharose beads. After washing and elution in sample buffer the bound proteins were separated on a gel, blotted, and detected by an anti-formin antibody ("Western blot"). Interestingly, formin cannot be co-precipitated with a α-SH3-antibody because the antibody appears to compete with the formin for the same binding site. The peptide to which the antibody α2–17 was raised competes for that site, too, so that formin is displaced from the binding ("+ peptide"). Source: modified after Uetz,P. *et al*. (1996) *J. Biol. Chem.*, 271, 33525–33530.

Depending on the antibodies only specific proteins bind to the matrix; other proteins can be easily removed by washing in buffer. Bound proteins can be removed and detected by boiling the beads in sample buffer and then separated on a protein gel. After staining the gel bands can be either cut out for analysis by mass spectrometry or the gel can be blotted for Western blot analysis.

Co-IPs are an important tool to verify PPIs, for example, Y2H interactions. The main limitation is that they require specific antibodies. If such antibodies are not available, proteins can be epitope-tagged and generic antibodies used to detect those tags. Many tags (or expression plasmids with such tags) and antibodies are commercially available (Table 16.3).

16.5 Far-Western

The idea behind the Western blot can also be used for the analysis of protein–protein interactions. Instead of detecting a protein on a membrane with antibodies, the blot can also be incubated with another proteins. Ideally this is done with an epitope-tagged protein. When this protein binds to protein on the membrane it can be detected by an antibody detecting this second protein. As in a Western blot, the second protein is visualized using ECL or a similar technology (Figure 16.13).

The method is pretty much the same as a regular Western blot, except for an extra incubation, which is why it is called "Far-Western". Far Westerns are used to confirm interactions that were found by a different method, such as Y2H interactions.

An interesting application of Far-Western-blotting is the mapping of interaction domains (or epitopes). This is achieved by partially digesting a protein and then separating the fragments on a gel. After blotting, the resulting membrane is incubated with the secondary protein, which

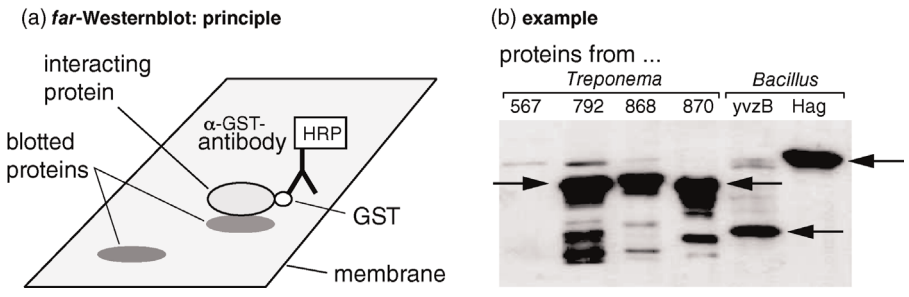


Figure 16.13 Far-Western blot: (a) principle (see text for details). (b) Example: The goal of this experiment was to show that the protein YviF (from *Bacillus subtilis*) binds to flagellar proteins from various bacteria. The flagellar proteins TP00792 (= FlaB2), TP0868 (= FlaB1), TP0870 (= FlaB3), and TP0567 (negative control), from *Treponema pallidum* (the causative agent of syphilis), are expressed alongside flagellar proteins yvzB and Hag from *Bacillus subtilis* in *Escherichia coli* cells. An extract from these cells is blotted onto a membrane. The membrane was then incubated with the GST fusion protein GST-YviF and the latter detected by an anti-GST antibody and ECL. HRP (horseradish peroxidase) is the enzyme that is coupled to the antibody and catalyzes a light reaction so that the protein can be detected by exposing the blot to an X-ray film. The Far Western shows that YviF binds to flagellar proteins FlaB1–3 and that this interaction is conserved across distantly related bacteria (namely, *Treponema* and *Bacillus*). Arrows indicate the flagellar proteins. All other bands are degradation products or unspecific interactions. Source: courtesy of Björn Titz and Seesandra V. Rajagopala.

Western Blotting, Section 5.3.3, 11.7

may bind only to some fragments but not others. From the size of the fragments it should become clear which fragment the protein is binding to.

16.6 Surface Plasmon Resonance Spectroscopy

Eva-Kathrin Ehmoser (Sinner)

University of Natural Resources and Life Sciences, Institute for Synthetic Bioarchitectures, Muthgasse 11-II, 1190 Vienna, Austria

Surface Plasmon Resonance, Section 5.3.3, 32.5.5

Surface plasmon resonance (SPR) spectroscopy is a method for detecting binding processes on surfaces. It originates from the phenomenon of surface plasmon waves and is today used to resolve subtle soft matter interactions at solid metal surfaces. SPR enables the measurement of binding constants in receptor–ligand assays including antigen-antibody interactions, DNA hybridization, and molecular interactions on functionalized surfaces and even whole cells. This technique offers the following advantages: sensitivity of detection in molecular dimensions, the use of low light intensities, label-free detection, the use of small sample volumes, absence of mechanical interaction with the sample, and reduced analysis time.

To plan and interpret SPR experiments meaningfully, it is useful to understand the underlying principle of the technique. Metal surfaces present electrons at the interface to the adjacent medium, which may be air or any other gas, water, a buffer solution, or an organic solvent. This adjacent medium is called dielectric. The electrons can be excited by directing light of certain wavelength and polarity onto the metal surfaces at a specific angle of incidence. The resulting oscillations of the electrons are called surface plasmons. Energy conservation dictates that the energy lost by the incident light striking the metal surface is transduced into propagation of the plasmons along the interface of the metal and the dielectric. The plasmon waves are of nanometer amplitudes, but are transmitted spatially over micrometer distances. The energy transduction is described as a plasmon coupling phenomenon or plasmon resonance.

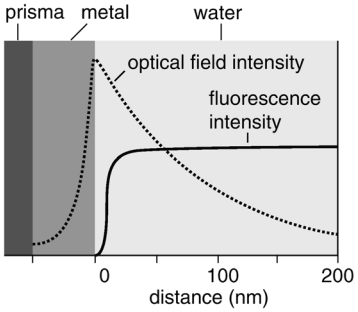


Figure 16.14 Surface plasmon resonance (SPR). Field intensity as a function of the distance from the surface.

The occurrence of resonance is governed by the angle of the incident light and the refractive index of the dielectric. Any change in the refractive index of the dielectric, such as that caused by adsorption of material onto the metal surface, would result in a change of the angle of incidence at which resonance occurs (Figure 16.14). Consequently, surface plasmon resonance is a surface-sensitive phenomenon. The outset of resonance is detected as a (steady) decrease in the intensity of the reflected light, or photon "sink", by a detection device, which may consist of a simple photo-cell or a CCD camera.

The most common experimental arrangement of a plasmon spectrometer is the so-called Kretschmann configuration, whereby a gold or silver layer is deposited onto a glass prism with a high refractive index. The resulting experimental set-up is depicted schematically in Figure 16.15.

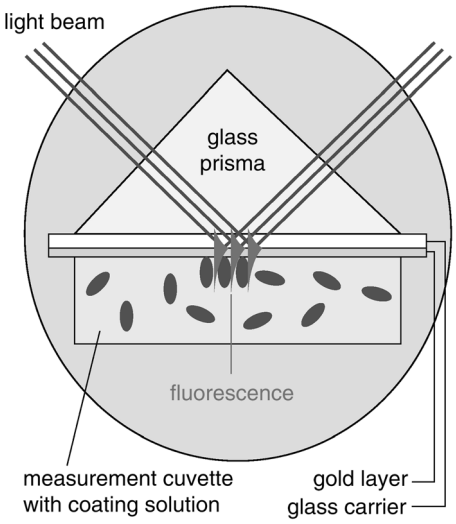


Figure 16.15 Basic setup of an SPR experiment.

Both, the prism and the metal surface can be finely tilted towards the light source, and by doing so the angle of light incidence can be precisely controlled. The angle of total reflection, beyond which the metal surface acts as a perfect mirror, is characteristic of the metal and the optical nature of the dielectric. Figure 16.16a shows typical outputs of a SPR experiment, wherein the reflected light detected by the photocell is plotted as a function of the angle of incidence, before and after the occurrence of a particular surface event (continuous and dotted lines, respectively). The curves display a minimum at the so-called coupling angle, where plasmon resonance occurs. The curves look very similar but appear somewhat shifted to one another; any change on the metal surface and hence, of the refractive index of the dielectric would alter the coupling angle and perturb the propagation of the plasmon wave. The change in the coupling angle is therefore detected as an horizontal shift of the SPR curve, which can in turn attributed to a particular binding process on the metal surface. SPR measurements are typically carried out at an angle of incidence where the change in the reflected light intensity, ΔR, before and after the surface event (R_1 and R_2, respectively) is particularly significant. Alternatively, ΔR can be recorded as a function of time in order to monitor surface transformations over time.

In a typical time-dependent binding assay involving two interacting molecules, one – the probe – would first be bound to the metal surface, forming a homogeneous layer. A buffer is then applied to the surface and a SPR measurement at a particular angle of incidence is conducted so as to obtain the baseline signal (D_1). The second molecule, or analyte, is then introduced. Any interaction between the analyte and the probe on the surface would be detected as an increase of the signal with respect to the baseline (Figure 16.16b). An example of such an assay would be that between an antibody and an antigen as probe and analyte, respectively.

SPR is an averaging method, which cannot differentiate between a homogeneous layer of molecules and a heterogeneous layer, e.g. one consisting of small molecules mixed with macromolecules. Consequently, it may be necessary to first purify and pre-characterized the material to be deposited on the metal prior to the SPR experiments. This is especially critical for biological samples.

Changes in refractive index may occur if the metal surface, especially plain gold, accumulates impurities. Non-specific binding of contaminants present in the analyte buffer may also occur. Changing the dielectric, such as switching from one buffer to another or even changing the buffer conditions, could also affect the coupling angle. Therefore it is essential that the SPR essay should be well designed to avoid signal interference caused by non-specific binding events or the mere exchange of buffer. As mentioned before, a pure, hydrophilic, gold surface would absorb virtually anything from the surrounding dielectric. One may hence think of substrate passivation as the way to render the gold surface apt for binding probe molecules. One means of doing this is to coat the gold with a layer of bacterial surface-layer (S-layer) proteins, to which the probe can then be immobilized via chemical conjugation. The S-layer proteins essentially block the gold surface from interacting with other molecules while preserving the plasmon propagation properties of the substrate, gold in this case. Alternatively, one could opt to use a non-fouling metal surface capable of plasmon resonance. Changes in buffer composition and conditions, such as pH and temperature, should also be minimized. Whenever possible, the use of ultrapure reagents and solvents should be encouraged to increase the signal-to-noise ratio.

Nonetheless, non-specific interactions are impossible to differentiate from specific interactions as they all lead to a change of refractive index and hence a measurable SPR signal. Even small temperature changes are reflected in the plasmon signal. As such, to make meaningful measurements of binding events at the metal surface, one needs to record a proper baseline to establish the background signal. Additionally, specific antibodies can be employed to "label" the actually present probe molecules bound to the surface. This can often be helpful, given the availability of respective antibodies. If performed properly, even subtle interactions between weakly interacting molecules can be detected.

SPR provide a measure of the so-called "optical thickness", a property greatly affected by the refractive index of the forming layer, which in turn changes as binding proceeds. However, it is possible to interpret optical thickness as physical thickness by the application of the "Fresnel" algorithm. Provided the refractive index of the layer is known and assumed to be a constant parameter, the algorithm allows the calculation of a layer thickness.

To date, SPR has proven a versatile and adaptable experimental approach for investigating molecular interactions in general and, particularly, biological samples from various sources. An

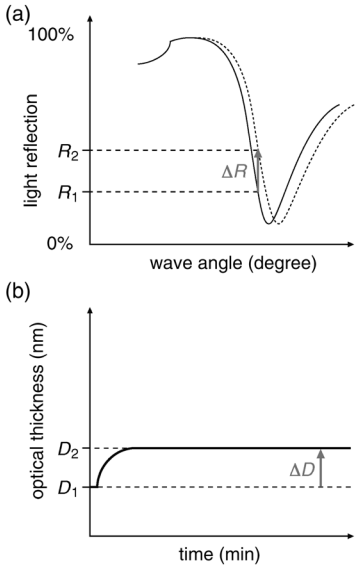


Figure 16.16 (a) Light reflection is dependent on its wave angle; (b) binding kinetics as time-dependent light intensity measurement.

interesting advancement is the modification of the metal surface with artificial membranes. These membranes can serve as scaffolds for the stabilization of hydrophobic molecules such as integral membrane proteins. This has allowed the use of membrane proteins in both their native conformation and orientation for the study of their interaction with ligands.

Although SPR is a label-free method, fluorescence enhancement by surface plasmons has been recently developed as an alternative method to detect molecular interactions. The incident light generates an evanescent field perpendicular to the direction of plasmon propagation, the intensity of which exponentially decays as a function of distance from the surface (Figure 16.14). This evanescent field can be used to excite fluorophores such as organic dyes or photostable quantum dots conjugated to analytes. In this case, the surface plasmons serve to enhance the fluorescence effect, making surface plasmon fluorescence spectroscopy (SPFS) possible. SPFS allows the plasmon waves to enhance sample fluorescence from a distance of up to 200 nm.

State-of-the-art SPR setups also employ microfluidic cells, in which the metal surface also acts as a seal. In this way, samples can be applied under microfluidic conditions, thus allowing users to exploit the various advantages of microfluidics, namely small sample size, laminar flow conditions, and automated fluid handling.

16.7 Fluorescence Resonance Energy Transfer (FRET)

Arnab Mukherjee

University of California, Santa Barbara, Santa Barbara, CA 93106, USA

16.7.1 Introduction

Förster resonance energy transfer (often referred to as fluorescence resonance energy transfer or FRET) describes the nonradiative transfer of fluorescence excitation energy between a pair of appropriately oriented fluorophores, typically referred to as the donor and the acceptor. FRET results in sensitized emission from the acceptor fluorophore and a concomitant quenching of donor fluorescence (Figure 16.17). Under these conditions, fluorescence emission from the acceptor fluorophore can be detected upon excitation at the donor wavelength. As a probe for

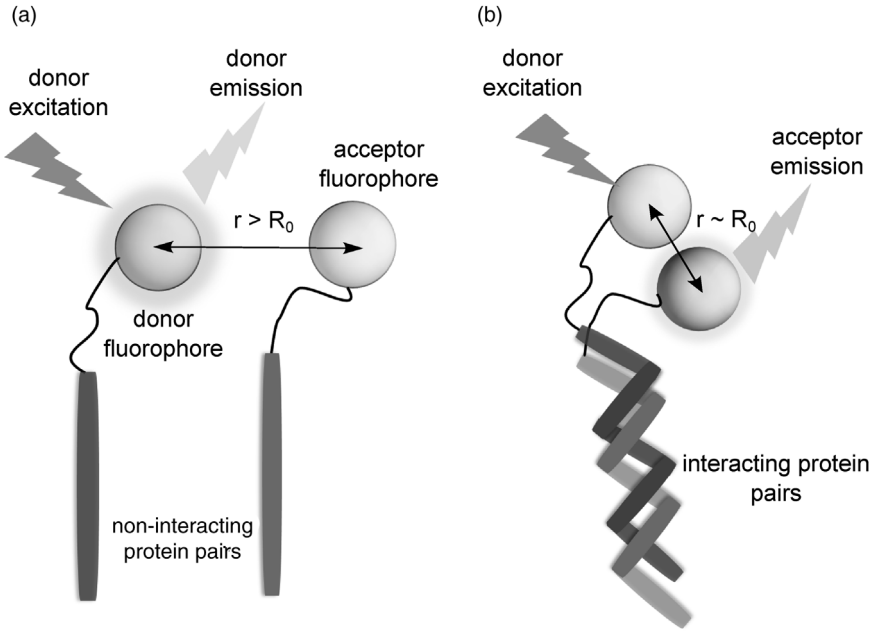


Figure 16.17 FRET is a sensitive probe for protein interactions. (a) Donor and acceptor fluorophores conjugated or translationally fused to a pair of non-interacting proteins do not demonstrate appreciable FRET as the donor–acceptor separation typically exceeds the Förster distance (R_0). (b) For interacting protein pairs, the donor and acceptor fluorophores are oriented in close proximity, which results in nonradiative transfer of energy from the excited state of the acceptor to the donor fluorophore, resulting in sensitized emission from the donor.

studying protein–protein interactions, FRET stands out in comparison to techniques such as yeast two hybrid, co-immunoprecipitation, and phage display in that FRET is uniquely suited for dynamic imaging of protein interactions in a biologically relevant cellular or sub-cellular context. By labeling proteins with fluorescent donor and acceptor dyes, FRET has been extensively used to elucidate the molecular underpinnings of chemotaxis, apoptosis, receptor oligomerization, circadian rhythms, host–pathogen interactions, and other cellular processes. In addition, FRET constitutes the physical basis of the vast majority of fluorescent biosensors designed for the detection of intracellular enzymatic activity, including kinases and proteases, ATP, oxygen, cell metabolites, trace metals, peroxides, and second messengers such as Ca^{2+} and cyclic AMP. In the following sections we present an overview of the theoretical basis of FRET and discuss the techniques that are most widely applied to analyze protein interactions via FRET. In addition, we describe emerging optical technologies for mapping protein interactions based on optogenetics and singlet oxygen sensitized quenching, which present viable alternatives to FRET-based interaction screens.

16.7.2 Key Physical Principles of FRET

The efficiency of FRET is primarily governed by the coupling between the dipole moments of the donor and acceptor fluorophores, which in turn is critically dependent on the donor–acceptor separation (r) and the overlap between the emission spectrum of the donor and the excitation spectrum of the acceptor. The FRET efficiency (E) is quantified as:

$$E = \frac{1}{1 + \left(\frac{r}{R_0}\right)^6} \tag{16.1}$$

where R_o (known as the Förster radius) is the distance at which the FRET efficiency is 50%. For a given donor–acceptor pair, the Förster radius can be calculated as:

$$R_0^6 = 0.211\kappa^2 n^{-4} J(\lambda) Q_D \left(\text{in units of } \mathring{A}\right) \tag{16.2}$$

where:

κ^2 describes the relative orientation between the donor emission and acceptor transition dipole moments;
n describes the refractive index of the environmental milieu of the donor–acceptor pair (1.33 for aqueous solutions);
Q_D is the quantum yield of the donor fluorophore;
$J(\lambda)$ quantifies the overlap between the donor emission spectrum and the acceptor absorption spectrum:

$$J(\lambda) = \int_0^\infty f(\lambda)\epsilon(\lambda)\lambda^4 \, d\lambda \tag{16.3}$$

where $f(\lambda)$ is the emission spectrum of the donor and $\varepsilon(\lambda)$ is the wavelength-dependent molar absorption coefficient of the acceptor fluorophore. The orientation factor (κ^2) is typically assumed to equal $^2/_3$ for fluorophores with rotational correlation times faster than the excited state lifetime. Owing to the strict dependence of FRET efficiency on the inverse sixth power of the donor–acceptor separation, FRET is highly sensitive to the distance between the coupled fluorophores. As is evident from Equation 16.1, FRET rapidly diminishes at distances $r > R_0$, which constitutes the physical basis for the application of FRET as a highly sensitive spectroscopic ruler for probing protein interactions and conformational changes with a resolution (1–80 nm) that is smaller than the optical diffraction limit of 200 nm.

16.7.3 Methods of FRET Measurements

FRET measurements can be made using live cells, purified proteins, or cell lysates. Methods for FRET measurement can be broadly classified into two categories – intensity-dependent and intensity-independent techniques. Intensity-dependent techniques measure sensitized emission

from the acceptor fluorophore or the concomitant quenching of the donor fluorescence upon efficient FRET. Intensity-independent techniques quantify the decrease in fluorescence lifetime or fluorescence anisotropy of the donor emission, which accompanies FRET. Intensity-based methods for FRET measurement require extensive correction for spectral crosstalk between the donor and acceptor fluorophores, which is discussed in Section 16.7.4. In contrast, although intensity-independent techniques are free from spectral crosstalk artifacts, fluorescence lifetime measurements necessitate elaborate and often expensive instrumentation, which may not be readily accessible to all research laboratories.

Instrumentation for Intensity-Based FRET Measurements Intensity-based FRET measurements are readily achieved using a spectrofluorimeter or a fluorescent microscope equipped with a suitable excitation source (e.g., 75 W or 100 W pressurized Hg or Xe lamps, He-Ne, Ar^+ ion, and diode lasers), a cooled electron multiplying CCD camera or a photomultiplier tube, a 60× or higher magnification high numerical aperture objective, and relevant optical filter sets. In particular, three filter sets are required for FRET detection – a donor filter set that selectively excites the donor fluorophores and detects emission from the excited donor, an acceptor filter set that excites the acceptor and selectively transmits the acceptor emission, and a FRET filter set that excites at the donor wavelength but transmits the sensitized acceptor emission, while ideally blocking out residual background emission from the donor. Suitably paired high quality excitation and emission bandpass filters are combined with longpass or quad-band dichroic beam splitters, which selectively transmit light at certain wavelengths while reflecting other wavelengths, thereby enabling spectral separation of excitation from emission fluorescence. Dichroic beam splitters can also be paired with a dual-view CCD camera that enables the simultaneous two-channel acquisition of both the donor and acceptor emission fluorescence. High quality optimized filter sets for FRET are available from several vendors (e.g., Semrock, Chroma) and are conveniently housed in rotating filter wheels on most modern microscopes. Advanced options include acousto-optic tunable filters (AOTF) and acousto-optic beam splitters (AOBS) that are a convenient alternative to planar filters and dichroic mirrors and obviate the need for carefully selecting and matching filter sets to the FRET pair.

FRET Estimation Based on Sensitized Acceptor Emission For an appropriately oriented donor–acceptor fluorophore pair, FRET results in fluorescence emission from the acceptor upon excitation at the donor wavelength. This provides a simple readout for FRET and can be quantified as the ratio of the fluorescence emission in the acceptor channel (S_F) to the quenched emission in the donor channel (S_D), which are respectively detected using a FRET filter set and a donor filter set (the subscripts F and D indicate the filter set used). For an interacting FRET pair, this ratio increases with efficient energy transfer between the fluorophores and the FRET efficiency (E) can be estimated as:

$$E = \frac{S_F}{(S_F + S_D)} \tag{16.4}$$

Although simple in principle, accurate detection and quantification of sensitized emission is rendered challenging by spectral bleed from two sources — (i) fluorescence emission from donor molecules that do not participate in FRET (it is difficult to precisely control the stoichiometry of donor and acceptor molecules to achieve a 1 : 1 interaction) and (ii) acceptor emission resulting from direct excitation at the donor wavelength vis-à-vis FRET (Figure 16.18). A widely used approach that accounts and corrects for spectral crosstalk involves conducting fluorescence measurements using "donor-only" and "acceptor-only" controls (e.g., cells individually expressing the proteins of interest labeled with the donor or the acceptor fluorophore) to estimate the relative contributions of the donor and the acceptor to spectral bleed-through. Specifically, the experimentally measured FRET signal (S_F) can be expressed as a linear combination of the actual FRET signal ($S_{F,corrected}$), fluorescence emission from unpaired donor molecules that do not participate in FRET (S'_D), and acceptor emission due to direct excitation at the donor wavelength (S'_A):

$$S_F = S_{F,corrected} + S'_D + S'_A \tag{16.5}$$

To deconvolute the FRET signal ($S_{F,corrected}$) from the S'_D and S'_A contributions, the following sets of fluorescence measurements are made under identical experimental conditions (excitation

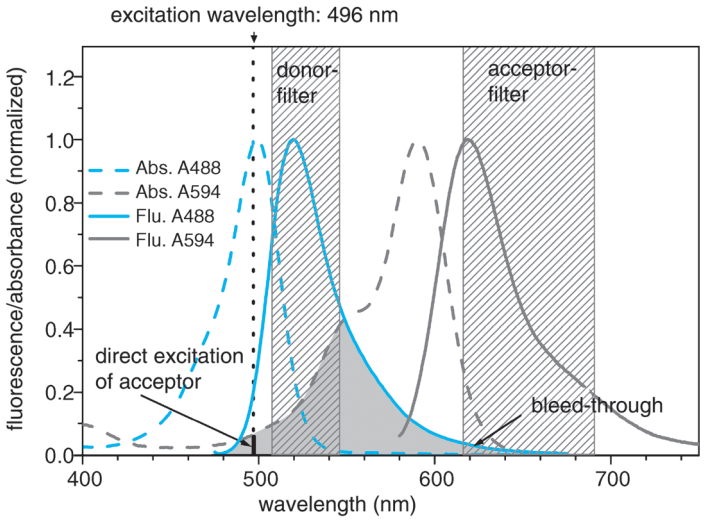


Figure 16.18 Microscopic or spectroscopic detection of FRET typically requires rigorous correction to eliminate spectral crosstalk resulting from direct excitation of acceptor emission at the donor wavelength as well as bleed-through of the donor emission into the acceptor emission filter window.

intensity, exposure time, microscope objective, etc.) using donor-only and acceptor-only controls as well as the donor–acceptor FRET pair:

1. fluorescence emission from the donor-only control measured using a donor filter set (D_D);
2. fluorescence emission from the donor-only control measured using the FRET filters (D_F);
3. fluorescence emission from the acceptor-only control measured using the acceptor filter set (A_A);
4. fluorescence emission from the acceptor-only control measured using the FRET filter set (A_F);
5. fluorescence emission from the donor–acceptor pair measured using the donor filter set (S_D);
6. fluorescence emission from the donor–acceptor pair measured using the acceptor filter set (S_A);

The FRET signal, $S_{F,corrected}$ can then be estimated as:

$$S_{F,corrected} = S_F - S_D\left(\frac{D_F}{D_D}\right) - S_A\left(\frac{A_F}{A_A}\right) \quad (16.6)$$

In the above equations, the subscript letters indicate the filter set employed (D, donor; A, acceptor; and F, FRET filter set), while *S, D,* and *A* represent experiments involving the donor–acceptor FRET pair, donor-only, and acceptor-only control, respectively. An alternative approach for deconvolution of the FRET signal involves spectral imaging and linear unmixing. In this approach, fluorescence emission from donor-only and acceptor-only controls as well as the donor–acceptor FRET pair are measured in multiple spectral channels using excitation at the donor wavelength. The relative contribution of a fluorophore's emission to each spectral channel is used to derive a spectral signature for that fluorophore, which is subsequently employed to deconvolute the FRET signal into the sensitized acceptor emission and quenched donor emission.

FRET Estimation using Acceptor Photobleaching Acceptor photobleaching represents a convenient approach for intensity-based FRET measurements that obviates the need for spectral deconvolution and spectral bleed-through correction. In this approach, the donor emission ($S_{D,pre\text{-}bleach}$) is initially measured for the FRET pair, which is followed by continuous excitation at the acceptor excitation wavelength in order to photobleach the acceptor fluorescence. Acceptor photobleaching dequenches the donor emission as FRET is abrogated and the enhanced donor emission ($S_{D,post\text{-}bleach}$) is recorded after the acceptor fluorescence is photobleached by at least 90%. Photobleaching conditions for the acceptor fluorescence need to be empirically determined using acceptor-only controls. For example, fluorescence emission from the yellow fluorescent protein variant, Venus, can be photobleached in fixed bacterial cells by

illumination for 2 min using a 488 nm laser line and an excitation power of 0.5 mW measured at the cover slip. Importantly, photobleaching of the acceptor must be achieved using a spectral bandwidth in which the donor excitation has negligible overlap in order to avoid photobleaching the donor fluorescence. FRET efficiency (E) can be computed as:

$$E = 1 - \frac{S_{\text{D,pre-bleach}}}{S_{\text{D,post-bleach}}} \tag{16.7}$$

Acceptor photo-quenching experiments, however, cannot be applied for dynamically imaging protein–protein interactions. Nonetheless, it is highly recommended to perform acceptor photobleaching experiments to validate FRET irrespective of the technique employed for FRET measurement.

FRET Based on Fluorescence Lifetime Measurements Fluorescence lifetime quantifies the amount of time that a fluorophore spends in its excited state before returning to the ground state. In the absence of FRET, the lifetime of a donor fluorophore (t_{D}) is governed by the rate of radiative decay (fluorescence emission) as well as nonradiative relaxation pathways, including internal conversion, intersystem crossing, dynamic, and static quenching. FRET provides an additional channel for fluorescence relaxation of the donor, which results in a decreased fluorescence lifetime of the excited state ($t_{\text{D,FRET}}$). In this case, FRET efficiency (E) can be estimated as:

$$E = 1 - \frac{\tau_{\text{D}}}{\tau_{\text{D,FRET}}} \tag{16.8}$$

Fluorescence lifetime imaging microscopy (FLIM) (Figure 16.19), however, is challenging to perform and requires elaborate instrumentation, which has prevented FLIM-FRET from being widely adopted for FRET compared to intensity-based techniques. Furthermore, several commonly used fluorescent donors (e.g., the cyan fluorescent protein or CFP) show multi-exponential lifetime decay, which complicates fluorescence lifetime based FRET measurements.

16.7.4 Fluorescent Probes for FRET

Although historically FRET has been widely studied and used in the context of small molecule organic fluorophores, the true potential of FRET as a powerful biological tool was fully realized with the advent of genetically encoded fluorophores based on the green fluorescent protein (GFP). Genetically encoded FRET probes enable facile labeling of proteins by constructing translational fusions, which circumvents the need to conjugate organic dyes to free amino or sulfhydryl groups in the protein *in vitro* and subsequently micro-inject the dye-labeled proteins for *in cellulo* experiments. Furthermore, genetically encoded FRET labels can be used to analyze protein interactions in relevant subcellular compartments such as the nucleus,

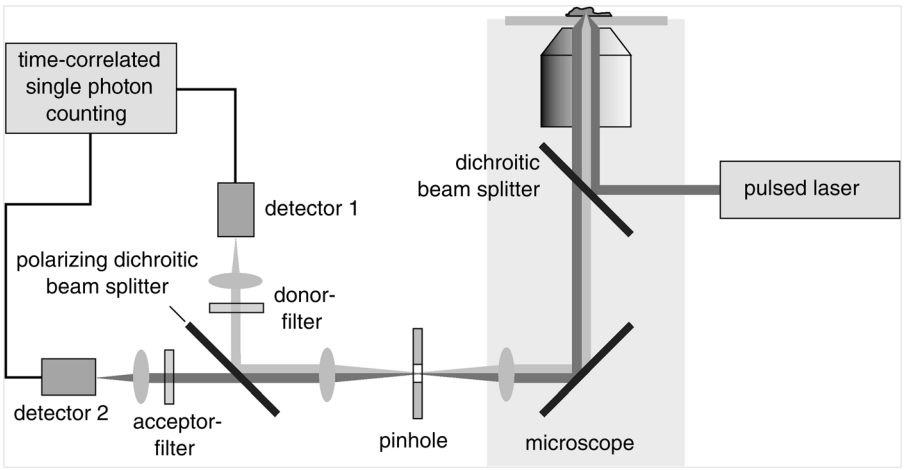


Figure 16.19 In a confocal construction for simultaneous lifetime and intensity measurement, a small volume of sample is illuminated with a pulsed laser and its fluorescent light is imaged onto a pinhole. The light passing through is separated by color or polarization, limited by spectral filters, and displayed on single-photon sensitive detectors. From the arrival time of the photons, the lifetime of the fluorescence can then be determined. To record images either the sample moves through the laser focus or laser focus moves through the sample with the help of moveable mirrors. Alternatively, for the detection of single molecules the focus can be fixed and the sample can be diluted to the point where the molecules move solely on the basis of diffusion individually through the focus.

mitochondria, and the cellular membrane. While small molecule dyes such as Cy3, Cy5, and the Alexa fluorophores are routinely used for *in vitro* FRET experiments (particularly for single molecule FRET imaging), genetically encoded FRET probes have become the mainstay for live cell imaging of dynamic protein interactions. Genetically encoded FRET probes, however, suffer from certain limitations, which must be kept in mind while designing experiments.

Genetically Encoded FRET Pairs — Caveats and Challenges An ideal genetically encoded FRET pair is characterized by a high quantum yield of the fluorophores, a large absorption cross section of the acceptor, single exponential decay of donor fluorescence, substantial overlap between donor emission and acceptor excitation, narrow spectral width and a large Stokes' shift to minimize spectral crosstalk, and appreciable photostability of both donor and acceptor. In addition, the translational fusions should be minimally perturbative to the native protein function. In practice, the aforementioned conditions are impossible to fulfil using a single pair of donor–acceptor fluorescent proteins and the optimal choice is largely guided by the application, imaging conditions, and experimental setup.

Cyan fluorescent protein (CFP, 436 nm excitation, 476 nm emission) and yellow fluorescent protein (YFP, 516 nm excitation, 529 nm emission) have been widely used as a genetically encoded donor–acceptor FRET pair. YFP is an attractive candidate as the acceptor fluorophore as it has a high quantum yield (0.61) and molar extinction coefficient ($84000\,M^{-1}\,cm^{-1}$). Although the emission spectrum of GFP (488 nm excitation, 509 nm emission) shows a greater overlap with the excitation spectrum of YFP compared to CFP, the small Stokes' shift of YFP (a general trait of most fluorescent proteins) renders it challenging to use GFP–YFP as a FRET pair owing to substantial donor emission bleed-through. Blue fluorescent protein (BFP, 383 nm excitation, 447 nm emission) is also used as a FRET pair with GFP as well as YFP. Although BFP minimizes spectral crosstalk compared to CFP, the low quantum yield and rapid photobleaching of BFP together with weak coupling between the BFP and YFP fluorophores complicates the application of BFP-YFP as a FRET pair. Red-emitting fluorescent proteins (e.g., mRFP1, mCherry) have been used as FRET acceptors in combination with CFP or GFP. Redshifted fluorescent proteins, however, are limited by their tardy maturation rates, sensitivity to photobleaching, and slow green-to-red maturation kinetics, which contributes to spectral bleed-through of the acceptor emission in the donor channel. Although several improved engineered variants of GFP are available for FRET (e.g., CyPet, Ypet, Cerulean, SCFP, SYFP, mCitrine, Venus), each genetically encoded FRET pair presents its own set of limitations and considerable caution must be exercised in carefully evaluating a particular FRET pair before using them for probing intermolecular interactions. In this regard, we highlight the following key caveats that need to be considered while using genetically encoded FRET pairs for imaging or quantifying protein interactions:

1. Donor and acceptor fluorescent proteins often show a tendency to dimerize (e.g., CyPet-Ypet), particularly under conditions of overexpression, which is common in transfected cells and which can result in false positive FRET signals. Cells co-transfected with donor and acceptor fluorescent proteins can serve as appropriate controls to check for spurious FRET signals resulting from interactions between donor and acceptor fluorescent proteins. In addition, several GFP variants can be monomerized through the inclusion of an A206K mutation, which can reduce the tendency of GFP-based FRET pairs to form heterodimers in cells.
2. Appropriate positive controls must be used to demonstrate FRET between the donor and acceptor fluorescent proteins. A widely adopted strategy to this end involves the construction of tandem donor–acceptor fusions separated by linkers of varying lengths. FRET should be observed for direct fusions and very short linker segments and the FRET efficiency should rapidly decrease with increasing linker length for non-interacting donor–acceptor pairs.
3. Fluorescent proteins (≈28 kDa) are considerably bulkier than small molecule based tags and may result in restricted rotational flexibility of the resulting translational fusions, which can reduce the FRET efficiency owing to improper coupling between the donor emission and acceptor transition dipole moments (the κ^2 factor in Equation 16.2). The rotational rigidity of a protein fused to a fluorophore can be estimated *ad hoc* using fluorescence anisotropy or polarization – anisotropy values <0.2 are typical for cyanine dye-conjugated proteins and

indicate considerable rotational flexibility. The use of flexible linkers consisting of glycine and serine residues connecting the protein of interest to donor or acceptor fluorescent protein is a widely used strategy to enhance the steric flexibility of fusion proteins. For particularly bulky fluorescent protein fusions, dipole alignment between the donor and acceptor fluorescent proteins can be tuned to identify a more favorable orientation by constructing circular permutations of the donor or the acceptor, which varies the orientation of the cyclic tripeptide chromophore within the β-barrel structure of GFP.

4. The Förster radius (R_0) for fluorescent protein FRET pairs is typically smaller than organic molecules and averages around 5 nm, which limits the sensitivity of FRET to distances of around 10 nm. Consequently, interactions between genetically labeled proteins may not result in appreciable FRET if the fluorescent protein labels on the interacting pairs are mutually separated by distances exceeding 10 nm, thereby resulting in false negatives. A viable strategy to address this issue involves the construction of translational fusions between the FRET pairs and both N- and C-termini (and possibly other positions within the protein) of the proteins of interest.
5. Translational fusions between the protein of interest and the donor or acceptor fluorescent protein must be verified to be minimally perturbative to the native protein function using functional assays of activity and cellular localization via immunofluorescence.

Bioluminescence Resonance Energy Transfer (BRET) A subtle, albeit highly useful modification to fluorescent protein-based FRET involves the substitution of the donor fluorophore with a bioluminescent protein derived from *Renila* or firefly luciferase, which is paired with YFP as an acceptor. The key advantage in employing luciferase as a bioluminescent donor is that it circumvents the need to use an excitation light source for exciting donor fluorescence emission, which in turn reduces the background from cellular autofluorescence and completely eliminates spectral bleed-through from direct acceptor excitation (S'_A term in Equation 16.5). Donor fluorescence is initiated by addition of a cell permeable substrate such as coelenterazine, which is oxidized by the luciferase. BRET can be quantified as the ratio of acceptor emission (at 530 nm for YFP) to donor emission (at 480 nm for luciferase) following addition of coelenterazine. The luminescence emission spectra of luciferases, however, are very broad and exhibit considerable overlap with the YFP emission spectrum, making spectral deconvolution difficult for imaging experiments. Nonetheless, BRET is highly suited for microplate based spectrofluorimetric assays for protein interactions and is an attractive option for probing protein interactions in photoresponsive systems such as retinal and plant cells.

16.7.5 Alternative Tools for Probing Protein–Protein Interactions: LINC and STET

More recently, alternative tools for imaging protein interactions have been reported, which leverage light-driven actuation of protein clustering or singlet oxygen generation to deliver an optical readout for interacting protein pairs. Although such tools do not directly exploit the resonance energy transfer mechanism, they can prove to be useful alternatives for supplementing as well as complementing the genetically encoded FRET tool box for imaging dynamic protein interactions in live cells.

Light Induced Co-clustering (LINC) Optogenetic actuation of protein interactions *in cellulo* is a promising approach for noninvasive, reversible, and precise spatio-temporal control of cellular activity, with unprecedented advantages over strategies based on chemo-genetic control. In the optogenetic approach, interactions between photoresponsive proteins (e.g., CRY2 and PHY domains) and their cognate partners (e.g., CIB and PIF domains) can be reversibly switched between bound and unbound states using pulses of light. Recently, the approach was extended for probing protein interactions based on light-driven clustering and co-localization of interacting proteins. In this approach, an engineered variant of a cryptochrome (CRY2olig) is optogenetically induced to cluster in response to blue light excitation (488 nm) into distinct punctae in the nucleus and cytosol. To probe for interacting protein pairs, a fluorescently tagged bait protein is translationally fused to CRY2olig and co-expressed with a prey protein that is labeled with a spectrally distinct fluorescent probe. Subsequently, the CRY2olig-tagged bait is clustered using a

short pulse (50 ms) of blue light. In the event that the prey and bait proteins interact, light-gated clustering of the bait protein results in co-clustering of the prey protein, which is manifested as a redistribution of the fluorescently tagged prey protein from a diffuse to a more punctate distribution that co-localizes with the bait clusters. Using this approach, multiple interacting partners of a bait protein can be identified by co-expressing prey proteins that have been labeled using fluorophores of different colors, which can be spectrally resolved. Light induced co-clustering (LINC), unlike FRET, is not sensitive to spectral crosstalk. Furthermore, as the clusters are reversible, with a half-life of 23 min, dynamic changes in protein interactions can be mapped by using light pulses separated by time intervals long enough to permit cluster dissociation. Although in its nascent stages, LINC could prove to be a highly useful alternative to FRET, particularly for cases where spectral bleed-through poses a significant challenge.

Singlet Oxygen Triplet Energy Transfer Singlet oxygen triplet energy transfer (STET) was recently introduced as a viable alternative to FRET for detecting long-range protein interactions in macromolecular complexes spanning several tens of nanometers. STET relies on the photosensitized generation of highly reactive singlet oxygen species from a donor fluorophore (known as the singlet oxygen generator (SOG)), which diffuses to a proximal acceptor fluorescent protein (known as the singlet oxygen sensitizer or SOS) and quenches acceptor fluorescence via oxidation of the acceptor chromophore. STET was realized using a genetically encoded SOG-SOS pair consisting of a fluorescent singlet oxygen generator from an *Arabidopsis* phototropin (known as miniSOG) and a biliverdin-binding infrared emitting fluorescent protein derived from a bacterial phytochrome (known as IFP1.4). Singlet oxygen production from miniSOG is initiated using blue light illumination (488 nm) while fluorescence quenching from IFP1.4 was detected using a 705/72 nm bandpass filter at the acceptor excitation wavelength (642 nm). As singlet oxygen-driven fluorescence quenching of IFP1.4 is rapid, STET is typically a diffusion limited process. Optimal spatial resolution for probing interactions using STET is therefore determined by the diffusion coefficient of singlet oxygen between the donor and acceptor fluorophores, which is several tens of nanometers in the cytosol. Although only a single SOG-SOS pair has been reported for STET, it is conceivable that optimized STET pairs will be engineered to enable versatile application of STET for detecting long-range macromolecular interactions that are outside the FRET limit.

16.8 Analytical Ultracentrifugation

Dieter Schubert[1], Klaus Richter[2], and Claus Urbanke[3]

[1]*Trabener Strasse 42, 14193 Berlin, Germany*
[2]*Technische Universität München, Department Chemie, 85748 München, Germany*
[3]*Hannover Medical School, Research Division of Structural Biochemistry, Carl-Neuberg-Straße 1, 30625 Hannover, Germany*

Analytical ultracentrifugation investigates the motion, size, or concentration distribution of biological or synthetic macromolecules in solution. Analytical ultracentrifugation has played an important role in biochemical research: Demonstrating the monodisperse nature and determining the molar masses of many proteins, nucleic acids, and supramolecular aggregates were landmarks in the development of biochemistry and molecular biology. During the 1970s and 1980s, many applications of analytical ultracentrifugation were taken over by other techniques, for example, electrophoretic methods or mass spectrometry to determine molar masses of proteins. However, the advancement of data processing techniques in the last two decades has enabled analytical ultracentrifugation to solve problems that were unfeasible previously. This is particularly true for the investigation of complex interactions of macromolecules and for the analysis of particle size distributions.

In contrast to other techniques like mass spectrometry or SDS-PAGE, analytical ultracentrifugation determines molar masses under native conditions. Especially in analyzing association equilibria this allows a correlation between protein function and molar mass distribution. Potential areas of application for modern analytical ultracentrifugation are protein–protein and protein-nucleic acid interactions. Typical problems in this field concern

self-association of proteins to dimeric, trimeric, and larger quaternary structures, the composition of protein complexes with different subunits, or that of a protein–nucleic acid complex. Since analytical ultracentrifugation measures accurate concentrations, quantitative conclusions on stoichiometry and equilibrium constants of the associations can be generated.

16.8.1 Principles of Instrumentation

Centrifugation, Section 1.5

In analytical ultracentrifugation, samples are observed in a running centrifuge; thus, the centrifuge has to have transparent sample cells and an optical system that allows quantitative measurements in them. Only one system used to be commercially available. It is based on the preparative ultracentrifuge *Beckman Optima XL*, which has been remodeled and contains an analytical addition that allows the spatial distribution of optical absorption to be measured at a given wavelength, throughout a double sector cell. The function is very similar to that of a double beam absorption spectrometer. The two sectors of the cell are filled with sample and solvent, respectively. An image of these sectors is projected by the optical system onto a plane just above a photomultiplier; the images are then scanned by a moving slit (Figure 16.20). After processing the intensity measurements, the output yields the absorption of the sample as a function of the radial position within the cell, $A(r)$. The radius scanned in this way extends over approximately 1 cm, and the volume of the sample typically lies between 50 and 450 µl. Additional optics offering other types of detection are available. Thus, interference optics measure changes in the refractive index n of the sample. Since n varies with the concentration of the dissolved macromolecules, their sedimentation can be observed by a radial analysis of $n(r)$. *Aviv Biomedical* supplies a fluorescence detector for the *Beckman Optima*. A 488 nm laser excites the chromophores in the solute molecules and the detector measures the reflected fluorescence at wavelengths above 505 nm. The laser moves radially along the sample cell and thus produces a concentration profile based on fluorescence intensities.

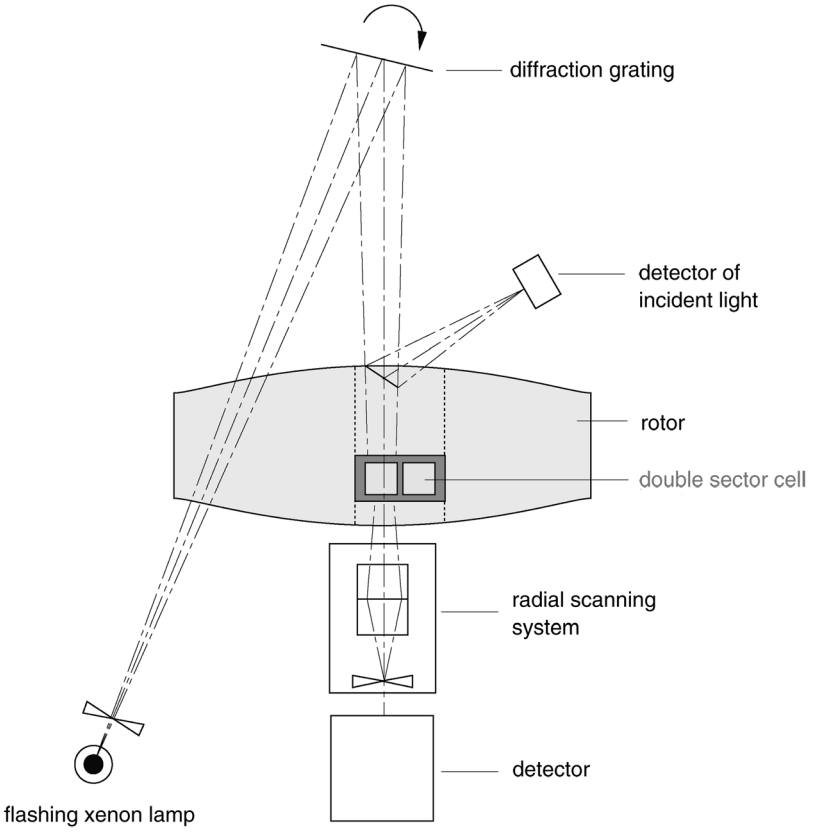


Figure 16.20 The photoelectric scanner in an analytical ultracentrifuge.

16.8.2 Basics of Centrifugation

According to Lambert–Beer's law, the local light absorption is directly proportional to the local concentration. Thus, measuring $A(r)$ yields the concentration distribution $c(r)$ within the sample cell. The information on the concentration is the central information analytical ultracentrifugation yields for the characterization of the sample. The same holds for the refractive indices $n(r)$ measured with the interference detector. Two laser beams, one through the sample and the other through the reference cell, form an interference pattern that is depicted by a CCD camera and converted into a fringe versus radius profile. This can be used directly to calculate the protein concentration since all proteins show nearly the same value of 3.3 fringes for a concentration of 1 mg ml^{-1} protein. Such a linear relationship between signal and concentration in principle holds for fluorescence detection, too. However, interactions between different molecules may change the fluorescence properties of the chromophores. An independent analysis of the quantum yield in different states of association may be necessary.

The concentration distribution (see Figure 16.21 below) is a function of experimental parameters (time of centrifugation, angular velocity of the rotor, temperature of the sample, distance from the center of rotation) and physical properties of the sedimenting macromolecules (molar mass, shape, density of the macromolecule) and the solvent (viscosity, density of the solvent) and thus allows the determination of several molecular properties. For all kinds of analytical centrifugation experiments the molar mass – the mass of the macromolecule corrected for its buoyancy (buoyant molar mass) – dominates the concentration profiles. In sedimentation velocity experiments, the angular velocity of the rotor is large enough to let all macromolecules finally sediment to the bottom of the cell. Kinetic properties of the macromolecule (sedimentation constant and diffusion and/or frictional coefficient) are important

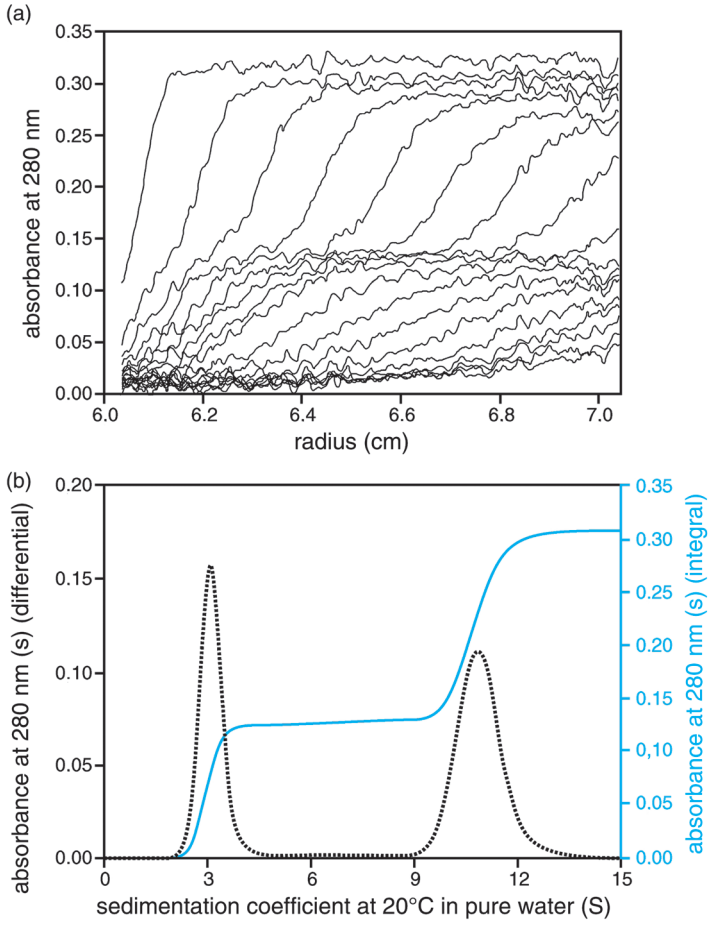


Figure 16.21 Sedimentation of a mixture of NAGK und PII: (a) sedimentation profiles; (b) *c(s)* distribution (- - - - differential; —— integral) calculated from (a) with SEDFIT. Source: After Maheswaran, M. *et al.* (2004) *J. Biol. Chem.*, 279, 55 202–55 210.

parameters. In sedimentation equilibrium experiments, lower angular velocities are used in such a way that sedimentation and diffusion balance each other. As a consequence, the macromolecules are dispersed between the bottom and meniscus in a defined manner. The concentration profile is solely determined by a single molecular parameter, the buoyant molar mass. Both types of experiments will now be dealt with in detail.

16.8.3 Sedimentation Velocity Experiments

Physical Principles A standard sedimentation velocity experiment starts with a homogeneously filled sample cell. Centrifugal, buoyancy, and frictional forces cause a uniform movement of the molecules, with large molecules in most cases running faster than small ones. A moving boundary is formed starting at the meniscus, and above this boundary the solution is free of sedimenting molecules (Figure 16.21). Analyzing the speed at which the boundary is moving yields the sedimentation coefficient, a characteristic property of the molecule. During sedimentation, the moving boundary is broadened by diffusion. A quantitative and complete description of this process is given by Lamm's differential equation, combining sedimentation and diffusion:

$$\frac{dc}{dt} = \frac{1}{r}\frac{d}{dr}\left[r\left(D\frac{dc}{dr} - s\omega^2 rc\right)\right] \tag{16.9}$$

where s and D are the sedimentation and diffusion coefficient, respectively, and c is the concentration of the macromolecule. By just observing the movement of the boundary over time, the sedimentation coefficient can even be determined manually. The sedimentation coefficient is a physical property of the molecule and can be related to its buoyant mass and its frictional coefficient, which in turn is related to the diffusion coefficient via Einstein's formula, finally resulting in what is called the "Svedberg equation:"

$$s = \frac{m(1-\bar{\nu}\rho)}{f} = \frac{M(1-\bar{\nu}\rho)D}{RT} \tag{16.10}$$

where: m and M are the mass of the single molecule and the molar mass of the species, respectively;
f is the frictional coefficient;
D the diffusion coefficient;
R is the gas constant;
T is the temperature (in K).

The term $\bar{\nu}$ is the partial specific volume of the molecules – to a good approximation this is the reciprocal density of the macromolecule – and ρ the density of the solution; $\bar{\nu}$ and ρ should be known independently of the sedimentation experiment. For proteins $\bar{\nu}$ can be calculated with sufficient accuracy from the amino acid composition. The density of the solvent can be calculated from its composition using standard tables or may be measured using, for example, a density bottle (pycnometer).

The sedimentation coefficient s has the unit of time and is usually given in "Svedberg units" S ($1\,S = 10^{-13}$ s). To make sedimentation coefficients comparable, they have to be corrected for the influences of solvent density ρ, solvent viscosity η, and temperature T of the experiment (index "T,S") with the standard condition being pure water at 20 °C (index "20 °C,W"):

$$s_{20\,°C,W} = s_{T,S}\frac{\eta_{T,S}}{\eta_{20\,°C,W}}\frac{1-\bar{\nu}\rho_{20\,°C,W}}{\bar{\nu}\rho_{T,S}} \tag{16.11}$$

Modern computer programs are now able to resolve the complicated relations between the measured concentration profiles and the molecular parameters s and D. Programs like ULTRASCAN, SEDFIT/SEDPHAT, or SEDANAL make a quantitative analysis of the properties of proteins as well as their association behavior possible. Numerical algorithms extract the parameters s and D from the measured concentration profiles by solving Lamm's differential

equation. These parameters are then used to calculate the molar masses of the species under study according to the Svedberg equation. Owing to the complex evaluation procedures and uncertainties, especially in the diffusion coefficient, confidence in the molar mass values is rather low (approximately 15%).

The same procedures can be used to investigate the quaternary structure of a protein and hetero-association between different proteins: If in a mixture of two or more macromolecules a sedimentation coefficient larger than that of any of its components is observed this can be interpreted as a physical proof of interaction between the macromolecules. In such a system of interacting molecules, the concentration of each species c_k is described by Lamm's differential equation with D_k and s_k unique for this species; the different differential equations then have to be combined using the law of mass action for the associations:

$$\frac{dc}{dt} = \sum_{k=1}^{n} \frac{dc_k}{dt} = \sum_{k=1}^{n} \frac{1}{r}\frac{d}{dr}\left[r\left(D_k \frac{dc_k}{dr} - s_k \omega^2 r c_k\right)\right] \tag{16.12}$$

Experimental Procedures Monomeric proteins larger than approximately 2 kDa can be characterized in sedimentation velocity experiments in terms of sedimentation and diffusion coefficients, leading to effective molar masses; additional information about the shape of the protein can be obtained via the frictional coefficient. With smaller proteins, the maximum speed of the rotor (60000 rpm) is insufficient to appreciably form sedimenting boundaries. Since differently sized proteins move with different velocities during centrifugation while all of them can be detected simultaneously, an analysis of protein mixtures and the characterization of the components of such mixtures is possible.

When analyzing the self-association of proteins, for example, in a monomer–dimer equilibrium, in principle a single sedimentation velocity experiment is sufficient to determine the dissociation constant of the complex. Since the masses of the monomer and the dimer are known, all other parameters can be obtained by computer-aided analysis of the data. However, in practice one would use different concentrations of the protein to obtain a "titration" curve of the monomer–dimer equilibrium, resulting in higher confidence and making additional information about the two states available. Computer analysis can be performed by SEDFIT/SEDPHAT or SEDANAL, which yield the appropriate constants.

Planning investigations on the interaction of two different macromolecules is far more complex. A simple association of n ligand molecules A to a single macromolecule B is described as follows. At a sufficient excess of free ligand A and fast equilibration (equilibration time less than 20 min), only two sedimenting boundaries are formed. The slowly moving boundary represents free molecules of A, and the faster sedimenting boundary represents all molecules of B regardless of whether they are free or complexed with A. With the sample cell homogeneously filled at the beginning, the sedimentation velocity experiment corresponds to an equilibrium dialysis where A is the component freely diffusible across the membrane, or to a size-exclusion chromatography experiment, where the smaller component is present in the running buffer (Hummel–Dreyer gel chromatography). Determining the interaction parameters (association constant and stoichiometry) now requires the concentrations of A in the free and bound state to be known. SEDFIT and ULTRASCAN can be used for this evaluation. This yields the concentration distribution of sedimentation constants in the experiment, with the diffusion of the particles being almost completely accounted for.

To actually determine the affinity of an interaction it is necessary to have all components, free and bound, at comparable concentrations. This is achieved whenever the initial concentrations are of the order of the dissociation equilibrium constant. Concentrations used for absorbance and interference detection are necessarily rather large (>μM range). Under these conditions, weak interactions can readily be analyzed, while for strong interactions stoichiometries, but not affinity constants, can be determined. However, the use of a fluorescence detector widens the concentration range, lowering the detection limit into the nM range and thus allowing strong interactions to be quantified. In addition, the amount of protein needed is greatly reduced. A

potential drawback for fluorescence detection is the need to modify the protein by chemically attaching a chromophore excitable at 488 nm. Several derivatives of fluorescein and intrinsically fluorescing proteins like YFP and GFP are available. There are other prerequisites when using sedimentation velocity to analyze interaction parameters. Above all, the difference in the sedimentation coefficient between the components and the complex should be as large as possible. When both molecules are detected by the optical system with similar sensitivity (absorption or interference) they should be of similar size. However, if the ligand is much smaller than the protein matrix, the ligand must be detectable since the mass increase of the protein will be too small to be detected. Dye markers in combination with fluorescence detection can be used to circumvent this problem. It is feasible to covalently couple small ligands or even small molecules with fluorescein to specifically detect their interaction with proteins. The same holds for macromolecular interactions, which can be analyzed by selectively labeling the smaller component. Under these conditions it is also possible to study large complexes built from several subunits by successive assembly, which is analyzed using specific labeling of a small component.

Example: Stoichiometry of the Interaction between *N*-Acetyl-L-glutamate-kinase and Signal Protein PII It has been shown that the activity of a key enzyme of arginine biosynthesis in the cyanobacterium *Synechococcus elongatus*, *N*-acetyl-L-glutamate-kinase (NAGK), was attenuated by the signal protein PII. NAGK is a hexamer ($M_W = 194.6$ kg mol^{-1}; $s_{20\,°C,W} = 8$ S), PII is a trimer ($M_W = 41.8$ kg mol^{-1}; $s_{20\,°C,W} = 3$ S). Figure 16.21 shows the sedimentation profile of a mixture of both proteins. Two sedimenting boundaries are observed. A $c(s)$ distribution calculated with the program SEDFIT/SEDPHAT shows two clearly separated peaks, with the faster one sedimenting at 10 S unequivocally demonstrating the interaction. The area under each $c(s)$ peak is proportional to the absorption and, thus, the concentration of the respective components. The absorption of the faster moving boundary is given by:

$$A_{fast} = A^{NAGK}_{total} + A^{PII}_{bound}$$

and that of the slower moving boundary by $A_{slow} = A^{PII}_{free}$. Knowing the initial concentrations and the extinction coefficients for both components, the concentrations of free and bound PII can readily be calculated. Figure 16.22 shows in a titration curve the concentration of free PII at different ratios of PII to NAGK; clearly a 1(trimer) : 1(hexamer) stoichiometry is demonstrated. The absence of any curvature in the titration curve shows the binding affinity to be too large ($K_{Diss} < 1$ μM) to determine an affinity constant at the concentrations given.

Another example of such titrations is provided by the interaction of E-cadherin, a surface protein of mammalian cells, with internalin, an adhesion protein from the bacterium *Listeria monocytogenes*. The interaction between these two proteins is a decisive step in *Listeria*

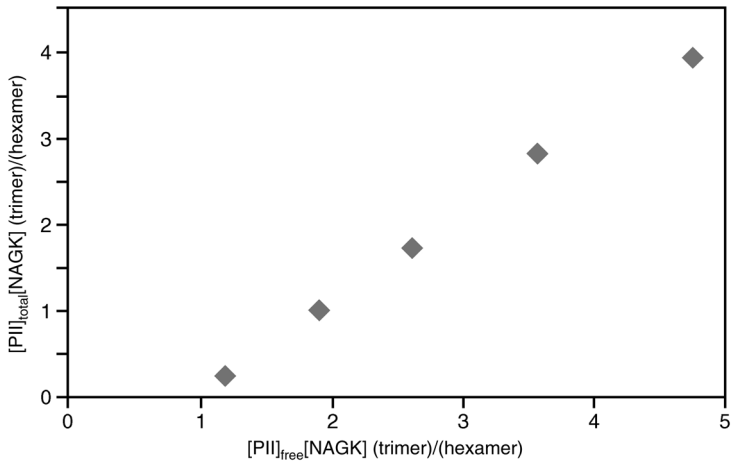


Figure 16.22 Titration of NAGK with PII protein (for details see text). Source: After Maheswaran, M. *et al.* (2004) *J. Biol. Chem.*, 279, 55 202–55 210.

infection. Sedimentation velocity experiments could show affinities in the range 10–100 μM. Stronger interactions, for example, even antigen–antibody interactions, can be analyzed with fluorescence detection.

16.8.4 Sedimentation–Diffusion Equilibrium Experiments

Physical Principles The influence of molecular shape in sedimentation velocity experiments has its pros and cons, owing to the large number of parameters that need to be determined. One way to elegantly circumvent the shape-related parameters *s, D,* and *f* is by the sedimentation–diffusion equilibrium method, which determines directly the molar mass of the species. It is one of the most accurate methods used to determine molar masses under native conditions. In its simplest application the method determines the molar mass of monodisperse molecules – from monomeric proteins up to supramolecular particles held together by non-covalent forces (e.g., viruses). Other applications deal with more complex associations (e.g., association equilibria). The sedimentation–diffusion equilibrium is a true thermodynamic equilibrium, implying that all other processes within the cell are at equilibrium, too. After equilibration at each point in the cell there is a different concentration of macromolecules (Figure 16.23) and the concentration gradient at any given position within the cell is a measure of the local weight average molar mass, which in turn is a function of the local state of the system.

The only technical difference between sedimentation velocity and sedimentation–diffusion equilibrium experiments is the lower speed of the rotor for the latter. In velocity experiments high rotor speeds are used to minimize the influence of back diffusion from the bottom of the cell on the sedimenting boundary. In equilibrium experiments this back diffusion is used deliberately to establish a sedimentation–diffusion equilibrium. In the evaluations, only data are used that are obtained after the net movement of the particles by sedimentation or diffusion has vanished and a stable equilibrium has been established. This takes between 24 h and several days, depending on sample and rotor speed. Under these conditions, $dc/dt = 0$ and Lamm's differential equation can be solved analytically; after combining it with the Svedberg equation the following function for the concentration distribution at equilibrium is obtained:

$$c(r) = c(r_0)e^{\frac{M(1-\bar{\nu}\rho)\omega^2\left(r^2-r_0^2\right)}{2RT}} \qquad (16.13)$$

where: r and r_0 are an arbitrary and a fixed position in the cell, respectively;
M is the molar mass of the macromolecule or the complex;
ω is the angular velocity of the rotor;
$\bar{\nu}$ is the partial specific volume of the macromolecule;
ρ is the density of the solution.

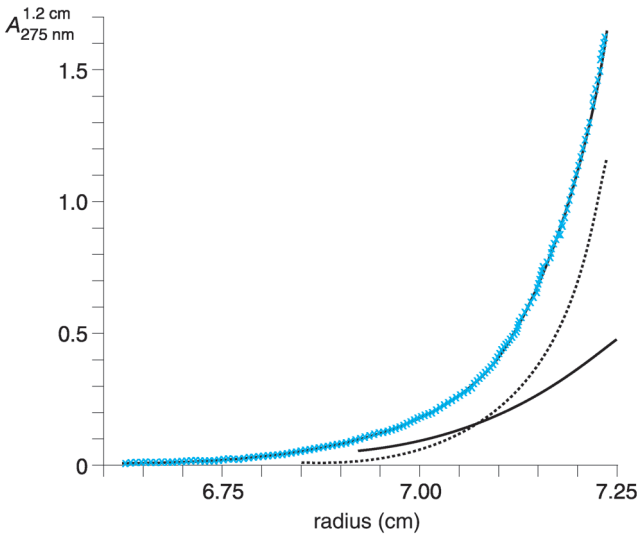


Figure 16.23 Self-association of a peptide (leucine-*zipper*) analyzed by sedimentation–diffusion equilibrium; $A(r)$ measured at $\lambda = 275$ nm (blue +), concentration distribution calculated for a monomer/dimer model (blue ——), concentration distribution calculated for the monomer (——) and dimer (·····). Rotor speed: 40 000 rpm. Source: Muhle-Goll, C. *et al.* (1994) *Biochemistry*, 33, 11296–11306. With permission © 1995, American Chemical Society .

The latter parameters have to be calculated or determined in different experiments, as described already in Section 16.7.3.

For heterologous mixtures and association equilibria, the concentration distribution for a mixture of n species can be described as the sum of the components c_i possibly linked by laws of mass action:

$$c(r) = \sum_{i=1}^{n} c_i(r_0) e^{\frac{M(1-\bar{v}\rho)\omega^2\left(r^2-r_0^2\right)}{2RT}} \tag{16.14}$$

With more complicated systems of several proteins or intricate assemblies, a thorough quantitative description becomes difficult. Although there are analytical expressions for the concentration distribution in analogy to Equations 16.9 and 16.10, the data measured do not yield enough information to separate the concentration distributions of the different species in a reliable manner. This increases the uncertainty of the interpretation for complex systems.

Experimental Procedures In protein analysis, frequently the focus is on the determination of the oligomeric state of the protein. Sedimentation–diffusion equilibrium experiments are the method of choice for such studies. Basically, a single concentration profile measured at a suitable concentration contains enough information to extract the association constant for a monomer–dimer equilibrium. Often, the molar masses of monomer (and dimer) are known a priori. In practice several sedimentation–diffusion equilibria at different concentrations are analyzed, with the concentration range covering monomer as well as dimer. Equation 16.14 is used for analysis. Analysis of larger oligomeric systems (trimers, tetramers, etc.) is also feasible with this method.

Interactions between different proteins can be analyzed with sedimentation–diffusion equilibrium experiments, too. However, experimental data only yield defined answers for systems containing not more than three to four concentration distributions $c_i(r)$. The associating system has to be quite simple. With large enough spectroscopic differences between the interacting partners, a simultaneous detection at several wavelengths or a combination with interference detection can aid in extracting species-specific concentration distributions (multi-wavelength detection). Labeling one of the molecular species with a dye may also help in the analysis of complex systems (bear in mind that the dye must not disturb the association behavior). Measurement at the absorption wavelength of the dye renders all non-labelled molecules and complexes invisible. Some information is lost; however, the information obtained can be interpreted more accurately. Similar arguments hold when using the fluorescence detection system, especially when labeling the smaller of the species. Adding to it the interaction partners leads to the formation of heavier complexes, and a titration up to saturation of the complex can be used to determine both stoichiometry and association constant.

Further Reading

Golemis, E.A. (ed.) (2006) *Protein-Protein Interactions: A Molecular Cloning Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor.

Piehler, J. (2005) New methodologies for measuring protein interactions in vivo and in vitro. *Curr. Opin. Struct. Biol.*, **15** (1), 4–14.

Wilson, W.D. (2002) Analyzing biomolecular interactions. *Science*, **295**, 2103–2104.

Section 16.1

Bartel, P. and Fields, S. (eds) (1997) *The Yeast Two-Hybrid System*, Oxford University Press, New York.

Battesti, A. and Bouveret, E. (2012) The bacterial two-hybrid system based on adenylate cyclase reconstitution in *Escherichia coli*. *Methods*, **58** (4), 325–334.

Caufield, J.H., Sahkarwalkar, N., and Uetz, P. (2012) A comparison and optimization of yeast two-hybrid systems. *Methods*, **58**, 317–324.

Fields, S. and Song, O. (1989) A novel genetic system to detect protein–protein interactions. *Nature*, **340**, 245–246.

MacDonald, P.N. (ed.) (2001) *Two-Hybrid Systems: Methods and Protocols*, Humana Press, Totowa.

Titz, B., Schlesner, M., and Uetz, P. (2004) What do we learn from high-throughput protein interaction data and networks? *Expert Rev. Proteomics*, **1** (1), 89–99.
Uetz, P. (2002) Two-hybrid arrays. *Curr. Opin. Chem. Biol.*, **6**, 57–62.

Section 16.2

Ben-Nissan, G. and Sharon, M. (2017) The application of ion-mobility mass spectrometry for structure. function investigation of protein complexes. *Curr. Opin. Chem*., **42**, 25–33.
Mehmood, S., Allison, TM., and Robinson, CV. (2015) Mass spectrometry of protein complexes: from origins to applications. *Annu. Rev. Phys*., **66**, 453–474.

Sections 16.3–16.5

Legrain, P. and Rain, J.C. (2014) Twenty years of protein interaction studies for biological function deciphering. *J. Proteomics*., **107**, 93–97.
Kimple, M.E., Brill A.L., and Pasker, R.L. (2013) Overview of affinity tags for protein purification. *Curr. Protoc. Protein. Sci*., **73**, Unit 9.9.
Jadwin, J.A., Mayer, B.J., and Machida, K. (2015) Detection and quantification of protein-protein interactions by far-western blotting. *Methods Mol Biol*., **1312**, 379–398.
Lin, J.S. and Lai, E.M. (2017) Protein-Protein Interactions: Co-Immunoprecipitation. *Methods Mol Biol*., **1615**, 211–219.

Section 16.6

Gordon, J.G. and Swalen, J.D. (1977) The effect of thin organic films on the surface plasma resonance on gold. *Optics Commun.*, **22** (3), 374–376.
Knoll, W. (1998) Interfaces and thin films as seen by bound electromagnetic waves. *Annu. Rev. Phys. Chem.*, **49**, 569–638.
Kretschmann, E. (1971) Die Bestimmung optischer Konstanten von Metallen durch Anregung von Oberflächenplasmaschwingungen, *Z. Phys.*, 313–324.
Otto, A. (1968) Excitation of nonradiative surface plasma waves in silver by the method of frustrated total reflection. *Z. Phys.*, **216**, 398–410.

Section 16.7

Lakowicz, J.R. (2006) *Principles of Fluorescence Spectroscopy*, Springer, New York.
Periasamy, A. (ed.) (2001) *Methods in Cellular Imaging*, Oxford University Press, Oxford.
van der Meer, B.W., Cooker, G., and Chen, S.Y. (1994) *Resonance Energy Transfer: Theory and Data*, VCH Publishers, New York.
Day, R.N. *et al.* (2001) Fluorescence resonance energy transfer microscopy of localized protein interactions in the living cell nucleus. *Methods*, **25**, 4–18.
Kenworthy, A.K. (2001) Imaging protein–protein interactions using fluorescence resonance energy transfer microscopy. *Methods*, **24**, 289–296.
Lo, T. *et al.* (2014) Singlet oxygen triplet energy transfer-based imaging technology for mapping protein–protein proximity in intact cells. *Nat. Commun.*, **5** (article 4072). Doi: 10.1038/ncomms5072.
Pfleger, K.D.G. *et al.* (2006) Illuminating insights into protein–protein interactions using bioluminescence resonance energy transfer (BRET). *Nat. Methods*, **3** (3), 165–174.
Piston, D.W. *et al.* (2007) Fluorescent protein FRET: the good, the bad and the ugly. *Trends Biochem. Sci.*, **32** (9), 407–414.
Sorkin, A. *et al.* (2000) Interaction of EGF receptor and Grb2 in living cells visualized by fluorescence resonance energy transfer (FRET) microscopy. *Curr. Biol.*, **10**, 1395–1398.
Taslimi, A. *et al.* (2014) An optimized optogenetic clustering tool for probing protein interaction and function. *Nat. Commun.*, **5**, article 4925. Doi: 10.1038/ncomms5925.
Truong, K. *et al.* (2001) The use of FRET imaging microscopy to detect protein–protein interactions and protein conformational changes *in vivo*. *Curr. Opin. Struct. Biol.*, **11**, 573–578.
Tsien, R.Y. *et al.* (1993) FRET for studying intracellular signalling. *Trends Cell Biol.*, **3** (7), 242–245.

Section 16.8

Kroe, R.R. and Laue, T.M. (2009) NUTS and BOLTS: applications of fluorescence-detected sedimentation. *Anal Biochem.*, **390**, 1–13.
Lebowitz, J., Lewis, M.S., and Schuck, P. (2002) Modern analytical ultracentrifugation in protein science: a tutorial review. *Protein Sci.*, **11**, 2067–2079.

Schubert, W.D., Urbanke, C., Ziehm, T., Beier, V., Machner, M.P., Domann, E., Wehland, J., Chakraborty, T., and Heinz, D.W. (2002) Structure of internalin, a major invasion protein of *Listeria monocytogenes,* in complex with its human receptor E-cadherin. *Cell*, **111**, 825–836.

Schuck, P. (2013) Analytical Ultracentrifugation as a Tool for Studying Protein Interactions. *Biophys Rev.*, **5** (2), 159–171.

Urbanke, C., Witte, G., and Curth, U. (2005) Sedimentation velocity method in the analytical ultracentrifuge for the study of protein-protein interactions, in *Protein–Ligand Interactions: Methods and Applications* (ed. G.U. Nienhaus), Methods Molecular Biology, vol. 305, Springer, pp. 101–114.

von Rückmann, B. and Schubert, D. (2002) The complex of band 3 protein of the human erythrocyte membrane and glyceraldehyde-3-phosphate dehydrogenase: stoichiometry and competition by aldolase. *Biochim. Biophys. Acta*, **1559**, 43–55.

Biosensors

17

Reinhard Renneberg,[1] Frieder Scheller,[2] and Aysu Yarman[3]

[1]Hong Kong University of Science and Technology (HKUST), Department of Chemistry, Clear Water Bay, Kowloon, Hong Kong
[2]University of Potsdam, Institute of Biochemistry and Biology, Karl-Liebknecht-Strasse 25-26, 14476 Potsdam, Germany
[3]Turkish-German University, Faculty of Science, Molecular Biotechnology, Sahinkaya Cad. No: 86, 34820 Beykoz No:86, Istanbul, Turkey

Biomolecules (enzymes, antibodies, nucleic acids) have been used for many years as analytical reagents. These biochemical recognition elements interact directly with the analytes to be detected and/or quantified. The formation of reaction products is then normally monitored with the help of various detection techniques (sensors, optical tests, Table 17.1).

In addition to the biochemical recognition elements on a molecular basis, membrane receptors, organelles, and whole living cells are also used. Biomimetic recognition elements mimic the function of biochemical "binding elements" and biocatalysts. These artificial systems are produced by chemical synthesis from monomeric building blocks. Amino acids or nucleotides are used for the production of aptamers or polymerizable monomers for molecular imprints (molecularly imprinted polymers, MIPs, Table 17.1).

Enzyme Activity Testing, Chapter 3

Immunological Techniques, Chapter 5

Nucleic Acid Analysis, Part IV

Table 17.1 Analytical configurations based on immobilized reagents.

Recognition element	Messenger/signal converter	Configuration
Biological recognition element	Photometer	
Enzyme	Interferometer	
Antibody	Refractometer	
Receptor	Fluorimeter	
Nucleic acid	Luminometer	
Organelle	Surface plasmon resonance	Test strips
Cell	Amperometric electrode	Analyzers
Biomimetic recognition element	Potentiometric electrode	Biochips
Aptamer	ISFETs	Biosensor
Ribozyme	Capacitive sensor	
Molecularly imprinted polymers	µTAS	
Synzyme	Thermistor	
Ionophore	Viscometer	
	Piezoelectric crystal	
	Magnetic toroids	

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

Following a trend in biotechnology, increasingly immobilized biomacromolecules (i.e., carrier-bound biomacromolecules) are used for analytics. The development moves towards integrated configurations:

- test strips
- biosensors
- biochips
- microanalysis systems (μTAS).

17.1 Dry Chemistry: Test Strips for Detecting and Monitoring Diabetes

Detecting and monitoring diabetes is one of the most important medical tasks. Around 7–8% of the population in developed countries on Earth are diabetics. *Diabetes mellitus* is a metabolic disease and characterized by a permanently increased blood glucose level. If the blood contains too much glucose, the kidneys cannot manage to filter it out and glucose is increasingly excreted via the urine.

The background reason of diabetes is the lack or disturbed action of the hormone insulin. Insulin is produced in the pancreas of mammals. It lowers the blood glucose level by supplying cells with the essential glucose. The normal blood glucose level is situated between 60 and 110 mg dl^{-1} and does not exceed 140 mg dl^{-1} after meals. Diabetics, however, show in fasting state levels of 126 mg dl^{-1} and above and after meals 200 mg dl^{-1} and higher.

Diabetics of type I lack insulin mainly due to destruction of β-cells of the pancreas. More than 95% of all human diabetes cases suffer, however, from diabetes type II. In these cases, the function and action of insulin is decreased. The body cannot compensate this even by producing more insulin. In recent years, a dramatic rise of diabetes has been observed. As mentioned, some 7–8% of the population in industrial countries are suffering. In rich Middle-East countries like Saudi Arabia as many as one-third of adults suffer from diabetes.

Historically, to detect diabetes, first of all urine was used, a non-invasive matrix. The success of the simple test strips for glucose detection in urine introduced in 1964 has led to other test strips: pH and protein (both in 1964), nitrite (1967); urobilinogen (1972), ketone (1973), bilirubin (1974) and erythrocytes (both 1974); and leucocytes (1982). Optical reading instruments have been developed in parallel for quantitation of qualitative test strips. As well as test strips with urine as matrix, serum, plasma, and whole blood test strips have also been developed. These are invasive tests and need a sterilized needle to take blood.

Nowadays, the diabetic or patient can determine "his or "her" blood glucose value by himself or herself. To manufacture multilayer test strips, technologies taken from color photography have been applied. Recent microprocessor technologies mainly simplified the use and control of these reading devices.

The "dry chemistry" methods used in most cases are quite well-known "wet chemistry" procedures. The water contained in the samples here dissolves the dry reagents. The conversion of the analyte on the test strips generates the measuring signal, for example, a colorless into a visible dye or an electrode-inactive into an active product. In most cases, a few microliters of the sample are sufficient and results are available within minutes.

Various types of immunoassays, mainly with enzyme labels (ELISAs) or colloidal gold, have been adapted for immuno-test-strips. These tests are competitive (for haptens) or sandwich configurations (real antigens). Most prominent is the pregnancy test for human choriogonadotropin.

17.2 Biosensors

17.2.1 Concept of Biosensors

Biosensors are based on the direct spatial combination of immobilized biochemical recognition elements and physicochemical signal transducers to quantify analytes in complex media.

At the beginning of 1960s, Leland Clark Jr., Gary Rechnitz, and George Guilbault in the USA positioned enzymes directly on electrodes. In 1962, Clark and Lyons first used oxygen electrodes

as transducers. To protect their systems from interfering substances, semipermeable outer membranes were used to cover the system. This direct integration of enzymes and transducers allowed the reuse of the precious enzymes. Clark called this configuration an "enzyme electrode" and applied for the first biosensor patent in history. However, the term "biosensor" was first used by Karl Camman in Germany in 1977. Enzyme electrodes are a subgroup of biosensors.

Today, the world market for blood glucose enzyme electrodes exceeds US$12.5 billion.

Following the use of enzymes, also intact living cells (cell sensors, microbial sensors), antibodies (immunosensors), and nucleic acids have been integrated in biosensors and been combined with novel transducers (Figure 17.1). Beginning in the mid-1970s, biosensors formed an independent branch of biotechnology at the edge to analytical chemistry. Milestones have been the start of the journal *Biosensors* in 1985 and the 1st World Congress on Biosensors in Singapore in 1990. Both were initiated by Elsevier.

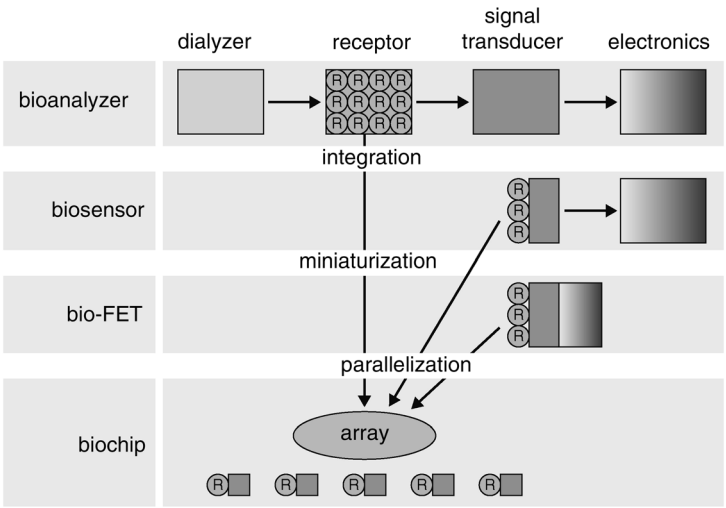


Figure 17.1 Developmental steps from analyzers via biosensors towards biochips.

17.2.2 Construction and Function of Biosensors

According to the IUPAC definition, "biosensors are devices which use a biological recognition element retained in direct spatial contact with transduction systems" (Thévenot, D.R., Tòth, K., Durst, R.A., and Wilson, G.S. (2001) Electrochemical biosensors: recommended definitions and classification. *Pure Appl. Chem.*, 71, 2333). The integration of biochemical recognition elements and transducers simplifies the analytical tool. This also would allow, additionally, a continuous measurement of the analyte by biosensors. All reagents have to be integrated into the sensor configuration.

In biosensors, the following processes are involved (Figure 17.2):

1. The recognition of the substance to be measured by the biological receptor.
2. The conversion of the physicochemical changes, caused by interaction with the analytes, into an electric signal.
3. Electronic signal amplification.

The biosensor scheme, developed by the authors of this chapter (Figure 17.1), shows clearly the essentials of the biosensor definition: the spatial integration of biological recognition and signal transduction.

From the sample (left-hand side in Figure 17.2), only low-molecular substances and molecular oxygen (blue in Figure) diffuse through the pores of a semipermeable outer membrane (middle of the figure) into the polyurethane gel. High-molecular substances, proteases and microbes cannot penetrate the membrane for steric reasons.

GOD only converts β-D-glucose with molecular oxygen into gluconolactone and H_2O_2. The immobilized high-molecular GOD cannot be washed out of the pores of the membrane. The product hydrogen peroxide is electrode-active at +600 mV versus Ag/AgCl. Its concentration

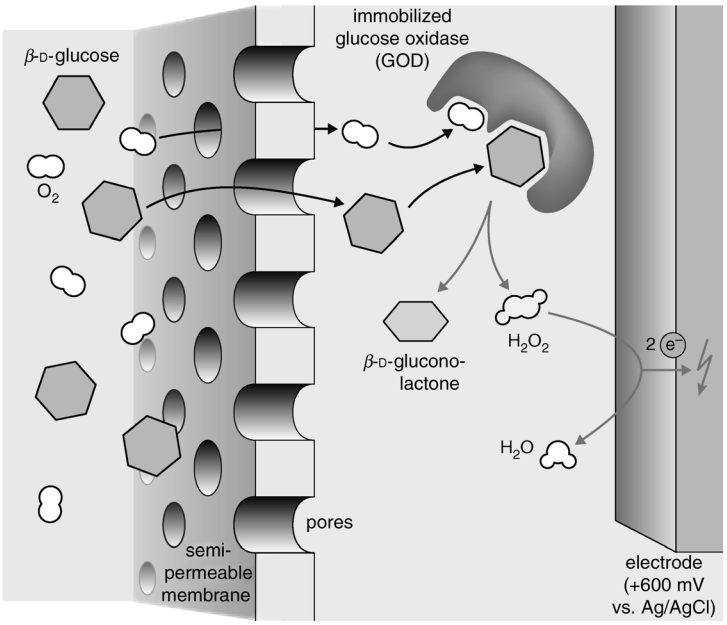


Figure 17.2 Reusable glucose biosensor: Clinical sensors, designed for thousands of measurements with one and the same immobilized enzymes, consist of an electrode covered with a thin enzyme membrane (here: glucose oxidase, GOD, is entrapped in a gel, e.g., polyurethane gel).

can be easily detected with electrodes. Finally, the concentration of glucose is directly proportional to the concentration of H_2O_2 and therefore proportional to the current at +600 mV versus Ag/AgCl.

For a measurement, the biosensor is dipped into the measuring solution containing the analyte. The biosensor is periodically calibrated with an equilibrated standard glucose solution. After measurement, the biosensor is rinsed with glucose-free buffer and can be reused again. One and the same biosensor can be reused 10 000–20 000 times. This is an amazing, cost-saving advantage, along with a high precision.

In analogy to affinity chromatography, based on specific recognition of ligands by bioreceptors, so-called "affinity sensors" have been developed: immobilized antibodies, tailor-made oligonucleotides, sugar-binding proteins (lectins), or hormone receptors are used for molecular recognition of antigens, complementary nucleic acids, glycoproteins, or hormones, respectively.

The physicochemical changes during complex formation, for example, the changes of layer thickness, refraction indexes, light absorption, and mass or charge distributions, can be monitored by optoelectronic, amperometric, and piezoelectric sensors, and also by field effect transistors (FETs) (Table 17.2).

In analogy to binding assays with antibodies or oligonucleotides, marker enzymes or fluorescent dyes are used for signal generation in so-called "indirect" sensors.

Biosensors based on molecular recognition of substrates and substrate conversion by biocatalysts are called "catalytic sensors." Here, the biosensor is regenerated during the catalytic analyte conversion. Continuous measurements are in principle possible as – after a short time lag – the velocity of the enzyme reaction follows the analyte concentration.

Table 17.2 Functional principles of biosensors.

Affinity sensor		Catalytic sensor	
	Analyte		Analyte
Dye	Protein		
Lectin	Saccharide, glycoprotein	Enzyme	Substrate
Enzyme	Inhibitor	Organelle	Cofactor
Apoenzyme	Prosthetic group	Microorganism	Inhibitor
Antibody	Antigen, hapten	Tissue slices	Activator
Receptor	Hormone		

Besides substrates, also cosubstrates (like NAD(P)H), activators, or inhibitors (e.g., the acetylcholine esterase inhibiting phosphorous organic pesticides) can be detected and determined with enzyme sensors.

A few years after the creation of enzyme sensors, intact living cells were already used for biosensing: microbial sensors, utilizing living bacteria or yeasts, are mainly applied nowadays for wastewater monitoring.

Generation of Enzyme Electrodes In living cells, chemical signals are mainly transduced by receptors. In enzyme sensors, however, low-molecular, diffusible reaction partners transfer the chemical signal from the enzyme layer(s) to the electrode (first generation). For oxidases, mainly the consumption of oxygen or the generation of H_2O_2 is monitored by amperometric electrodes, for dehydrogenases it is the conversion of cosubstrates like NAD(P)H, and for peroxidases it is the H_2O_2 consumption (Figure 17.2). The change of pH values due to substrate conversion by hydrolytic enzymes is monitored by potentiometric electrodes.

Several oxidases can use artificial low-molecular redox substances, so-called mediators (Figure 17.3). This makes measurements independent of both oxygen concentration and disturbing interferences (oxidizable substances, e.g., ascorbic acid in samples).The integration of mediator molecules in the sensor (e.g., in carbon paste or as redox polymers on the electrode surfaces) leads to reagentless measurements and biosensors (second generation).

The third-generation biosensors use the direct (mediator free) electron-transfer between enzymes and the redox electrodes. Most oxidoreductases hide the redox-active groups inside the molecules, for example, the FAD prosthetic groups of glucose and lactate oxidases, where they are well protected. Therefore, the direct communication of these so-called "intrinsic" enzymes with redox-electrodes is sterically hindered or impossible. In contrast, "extrinsic" redox enzymes (e.g., the dehydrogenases for cellobiose, glycolate, fructose, and methylamine conversion) possess prosthetic groups accessible for heterogeneous-transfer. After adding proper substrates, catalytic currents can be measured.

In addition, for cytochrome *c,* laccase, and also peroxidase a direct electron-transfer was realized. To transfer redox equivalents also from electrode-inactive redox enzymes, mediators (e.g., ferrocene) have been covalently bound either directly to the enzyme protein or via mobile spacers (molecular wires made out of redox active polymers). Metallic nanoparticles and carbon nanotubes can also provide effective electron-transfer.

To check for fitness, lactate biosensors are on the medical device market that use lactate oxidase (LOD) instead of GOD. LOD converts lactate into pyruvate using its redox-active centers ($FAD/FADH_2$), under consumption of oxygen. Oxygen is the natural electron-acceptor

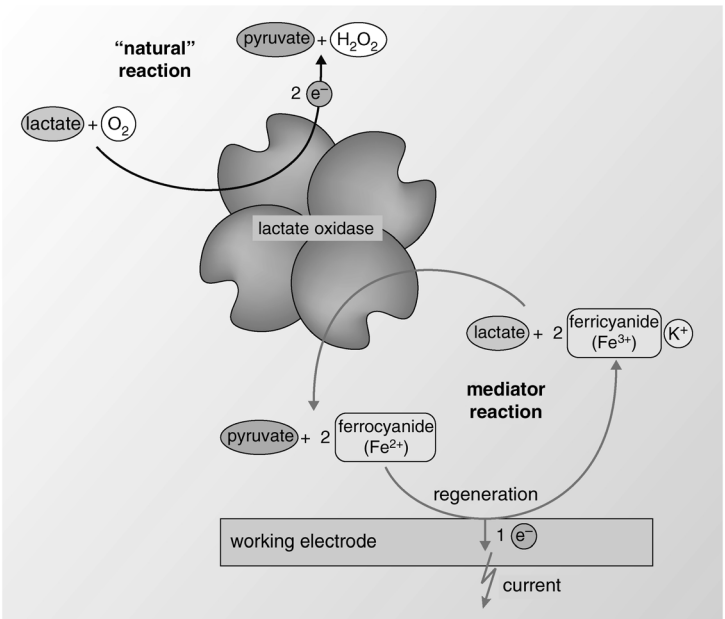


Figure 17.3 Second generation enzyme electrode for lactate.
In analogy to the application of glucose oxidase in blood glucose sensors the enzyme lactate oxidase is used in lactate sensors which are applied in sports medicine. Lactate oxidase catalyzes the conversion of its substrate to pyruvate by oxygen. The catalytic center of the enzyme contains flavine adenine dinucleotide ($FAD/FADH_2$) which is reduced by lactate and re-oxidized by ambient oxygen. In this reaction oxygen is consumed and hydrogen peroxide is produced (upper left). Because for the oxidation of one lactate one oxygen is consumed, the low oxygen concentration is not sufficient for the complete conversion of the lactate which has a considerably higher concentration in blood. In order to overcome this limitation the lactate sensors for undiluted blood contain an artificial electron acceptor for example ferricyanide. It substitutes oxygen in the enzyme catatalyzed conversion of lactate to pyruvate and is reduced from its ferri form (Fe^{3+}) to ferrocyanide (Fe^{2+}). The ferrocyanide is regenerated to ferricyanide at the working electrode of the sensor chip. This electrode process generates the measuring current which is linearly dependent on the lactate concentration in the sample. In addition falsification of the lactate value by ascorbic acid or paracetamol are prevented because the electrode works at a lower potential than for the detection of peroxide (lower right).

of LOD, taking up the electrons from lactate and reducing molecular oxygen to H_2O_2 (Figure 17.3 upper part). Oxygen is only soluble in water in low concentration. It would limit the reaction, because lactate is present in whole blood in much higher concentrations. To solve this problem in a smart way, for one-way lactate biosensors an artificial electron-acceptor is used in excess concentration. LOD now oxidizes lactate to pyruvate and transfers in parallel the released electrons to the ferricyanide (Fe^{3+}), which is reduced to ferrocyanide (Fe^{2+}). By the transfer of one electron to the working electrode of the sensor chip, ferric cyanide is regenerated and can be reused. Due to this electron transfer, a current flow arises that is proportional to lactate concentration (Figure 17.3 lower part).

Coupled Enzyme Reactions in Biosensors The coupling of different enzyme reactions widens the analyte range and leads to an internal chemical signal transformation in biosensors (Figure 17.4).

Sequence and Competition Some co-substrates and products of enzymatic reactions are not detectable with current transducers as they are not electrochemically active. This limits indeed the number of substances that are detectable with mono-enzyme sensors. To generate the desired electrode-active products, additional enzymes have to be coupled in sequential mode following the conversion of the target analyte. Enzyme sequence electrodes have been described for cholesterol esters, fatty acids, amides, and oligosugars. With multi-enzyme sensors the total sum of analyte concentrations can be determined if a complete conversion is supported in the enzyme membrane and the diffusion velocities are comparable.

The signal results from the *addition* of two substrate concentrations (Figure 17.4). The *competition* of two enzymes for one and the same substrate in the presence of the cofactor is used in enzyme sensors for adenosine triphosphate (ATP) and nicotine adenine dinucleotide (NAD^+). GOD is co-immobilized with hexokinase or glucose dehydrogenase, respectively. If only glucose is present in the measuring solution, it alone will be converted by GOD. After addition of ATP or NAD^+, however, a part of the glucose is now consumed by the competing enzymes. The signal corresponds to a *subtraction* of the signals of both separate biosensors: for glucose and of the biosensor for the corresponding cofactor ATP or NAD^+.

Anti-interference Principle With an increasing number of enzymes in an enzyme sensor, the selectivity is decreased, as the substrate(s) of each additional enzyme may cause signals. In addition, some substances in the sample may cause disturbances, if they are converted chemically at the transducer. To eliminate such interfering substances, enzymatic anti-interference layers have been integrated into biosensors. For example, for glucose detection in urine the interference by ascorbic acid can be eliminated by adding to the enzyme membrane a layer of ascorbate oxidase (or laccase). Both enzymes circumvent the anodic oxidation of ascorbic acid by enzymatic conversion into the electrode-inactive dehydroascorbic acid.

Enzymatic Analyte Cycles The lower detection limit of conventional enzyme electrodes is situated around 1 µmol l^{-1}. After dilution, the sensitivity for most hormones, pharma products, and some metabolites is not sufficient. The sensitivity, however, can be multiplied, if the analyte is (re)cycled between two enzymes, that is, it is converted and regenerated back many times (Figure 17.4). In this way, the "sensitivity threshold," caused by diffusion, can be overturned. At least one of the two reactions of the cycle should involve an electrode-active substance (e.g., oxygen). By enzymatic amplification a much higher oxygen consumption is

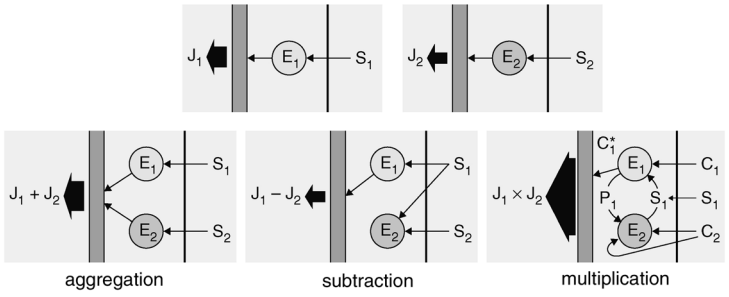


Figure 17.4 Coupled enzyme reactions in biosensors.

performed than with a single simple analyte conversion. An amplification of 10 000 could be reached for detection of catecholamines. This extends the measuring range into sub-nanomolar concentrations.

17.2.3 Cell Sensors

In 1987, ten years after the first enzyme electrodes, in 1977, Isao Karube and Shuichi Suzuki constructed the first cell electrodes. They applied bacteria, yeasts, but later also mammalian cells (e.g., cancer cells and neurons). Microbial cells have been finally convincing for their stability. Instead of single enzymes, like the glucose and lactate sensors, complex enzyme systems were used: intact living cells. The respiration rate is measured. This is a group parameter, as all relevant influences from the sample are monitored, for example:

- the nutrient yield in fermentation media;
- biodegradable substances in wastewater;
- mutagenicity of substances;
- cancerostatic effect of drugs.

Cell sensors allow therefore the quantification of biological effects when the respiration rates of living cells are measured. Cell sensors allow quantification of biological actions, which is impossible with chemical methods. Cell chips characterize the physiological state of cells. They measure, besides cell respiration, also the acidification in the cell environment with the help of pH-FETs, the consumption of nutrients (e.g., glucose and amino acids), the lactate formation, and the viability with the impedance of interdigital electrodes.

Microbial Sensors, Biochemical Oxygen Demand of Wastewater When wastewater is delivered to lakes, rivers, and the sea, the concentration of dissolved oxygen (DO) is decreased dramatically. Aerobic bacteria and fungi, however, need oxygen for the biodegradation of organic waste. Wastewater treatment facilities are biofactories for the production of clean water. They need additional oxygen to fulfill their function. As in an aquarium, oxygen from the air is pumped into water.

The procedure of biochemical oxygen demand in 5 days (BOD5) was introduced in England as early as 1896 to determine the organic load of wastewater. Even today, BOD5 is a standard procedure, allowing the determination of easily biodegradable substances of the total organic load. The oxygen consumption by heterotrophous microbes at 20 °C is measured for 5 days (BOD5): The water sample is diluted, saturated with air, and microbial seeds are added. The seeds are normally a mixed culture of wastewater microbes for inoculation. The BOD5 is important for comparison reasons for wastewater – the financial charges made for wastewater are calculated according to it. It gives, however, no hint at non-biodegradable substances. The long duration of 5 days is also a disadvantage. It does not allow on-line monitoring and feedback regulation of wastewater facilities. In contrast, microbial biosensors provide results within 5 min!

Living microbes immobilized on sensors, for example, yeasts like *Trichosporon cutaneum* and *Arxula adeninivorans* allow us to determine in time the organic load of wastewaters. The halophilic (salt-tolerant) yeast *Arxula* is best suited for wastewaters in coastal tropical countries. In Hong Kong, for example, toilets are flushed with sea water. The wastewater has therefore a high salt content, which inactivates typical salt-sensitive microbes, but not *Arxula*.

As noted, the BOD5 test takes 5 days while the BOD biosensor, in contrast, takes only 5 min. Living yeast cells are immobilized in a polymer gel (as in reusable glucose biosensors) and the membrane is mounted on an oxygen electrode. The respiration rate (slope of oxygen consumption) of "starved" cells is then measured. If a clean water sample is added (no biodegradable substances), the "hibernating" yeast does not pick up additional oxygen and is in a stand-by mode. But as soon as a sample contains "food", the cells are activated, take it up and consume it using molecular oxygen. The consumed oxygen is proportional to the amount of biodegradable substances. The biosensor is calibrated with a standard of glucose and L-glutamate.

Microbial sensors are ideal for monitoring wastewater facilities: The microbial biosensor shows how polluted the incoming wastewater is and regulates in feedback mode the aeration pumps for the activated sludge process. In this way, a considerable amount of energy can be

saved. Another microbial biosensor at the exit shows the actual degree of water cleaning in the facility.

17.2.4 Immunosensors

Surface Plasmon Resonance, Section 5.3.3, 16.6, 32.5.5

Besides immuno-test-strips and ELISAs, biosensors based on antibodies (immunosensors) have also been developed. Immunosensors detect the recognition and binding of antigens by antibodies. As with other biosensors, the recognition elements (antibodies or antigens/haptens) are immobilized on transducer surfaces (Figure 17.5).

Label-free "direct" immunosensors transduce the antibody–antigen interaction directly into color changes or signals of surface plasmon resonance (SPR) or piezoelectric crystals. "Indirect" immunosensors use additional labels (enzymes, fluorophores) to indicate the interaction.

Recently, easy to handle commercial devices have become available, mainly the Biacore type SPR and also fiber-optical devices. They are easier to handle and lighter than the original prototypes.

Indeed, today's fully automated fiber-optical sensors are much lighter and more reliable than the original 70 kg weight prototypes. Now, eight substances/microbes can be monitored simultaneously and the device is integrated with an airborne germ collector and carried as bag-pack. Another version, carried by a small unmanned drone, can identify bacteria during its flight.

Most of these immunosensors use the following principle: The catcher antibodies are bound via avidin/biotin to glass surfaces, which are a fiber-optics or a simple glass slide. A laser beam is then guided through the glass to its antibody-bearing surface. At the interface between glass and liquid phase, the two media have two different refractive indices. If the beam hits the surface at an angle that is smaller than the critical angle, total internal reflection (TIR) of the beam takes place (Figure 17.6). TIR generates an evanescent wave on the glass surface. This

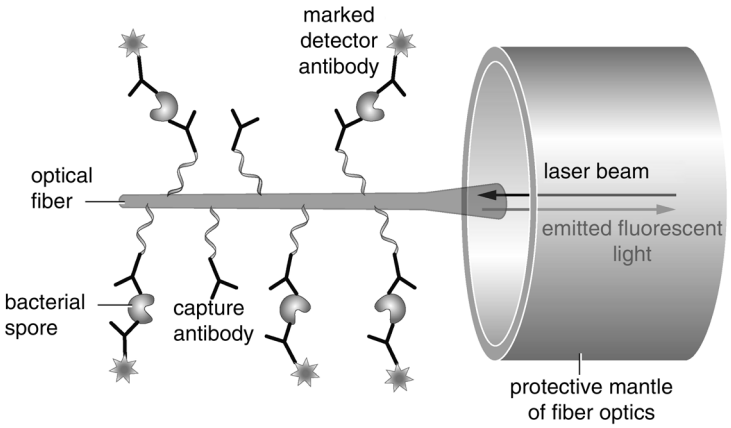


Figure 17.5 Immunosensor using the effect of evanescent waves to excite fluorescent markers bound to detector antibodies (for details, see text).

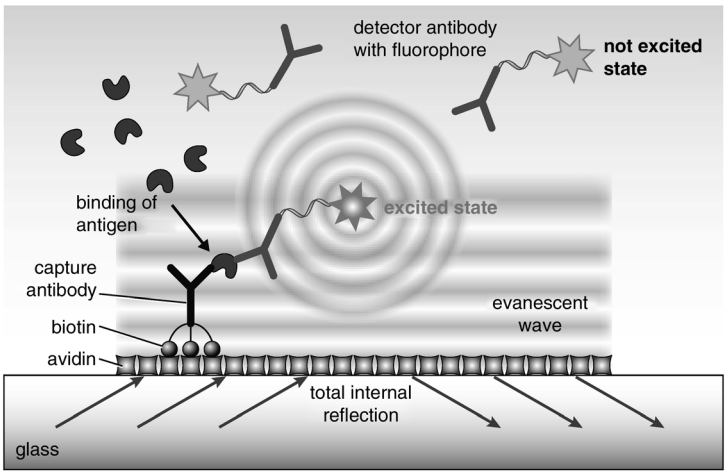


Figure 17.6 Principle of a fiber-optic biosensor for the detection of bacterial spores. The labeled detector antibodies are excited by the evanescent wave.

wave penetrates into the fluid for around 100 nm, almost the radius of the antibody–antigen interaction.

Two different antibodies are used for detection. First, an antibody for binding the antigens (catcher or coating antibody), which generates an evanescent wave on the glass surface. This wave penetrates the fluid layer for around 100 nm (almost the radius of the antibody–antigen complex); second, fluorescence-labeled detector antibodies are bound to the antigen. They form a sandwich as in ELISAs, where enzyme labels are used. A sandwich construct is around 30–50 nm in diameter. Therefore, the fluorescent markers are in exactly the area of the excitation energy of the evanescent wave!

The marker is excited and consequently emits light. This means that the antigen is recognized and bound! Unbound labeled detector antibodies are simply not excited because they are mostly located beyond the evanescent wave.

Finally, the fluorescent signal is filtered and amplified. These immunosensors work in the ppb region (parts per billion). This corresponds to a spoonful of salt in a swimming pool of Olympic size. Meanwhile, array immunosensors have been developed to monitor several analytes in parallel. These systems consist of a multitude of different antibodies in clearly separated areas (fields, dots) on flat carrier material. Sample molecules and fluorescent labels in some areas yield a signal, in other areas they do not. Like in DNA chips, the identity of a sample molecule is determined by the location of the fluorescent area. The intensity of the signal reveals the amount of the target molecule in the measured sample.

Novel antibody based biosensors are highly sensitive and are used to monitor pesticides in agriculture, toxins and pathogens in homogenized food, disease markers in clinical fluids, and biowarfare agents in air and water.

17.3 Biomimetic Sensors

Evolution has brought about biopolymers on the basis of amino acids and nucleotides showing high chemical selectivity and catalytic power. Molecular recognition and catalytic conversion of the target molecules by antibodies and enzymes is accomplished in so called epitopes or catalytic centres that typically consist of 10–15 amino acids. Nucleic acids – which are made up of only four different nucleotides – bind not only complimentary single stranded nucleic acids by base pairing (hybridization) but also interact highly specifically with proteins (e.g., transcription factors) and low-molecular weight molecules and even with ions.

As mimics of biological macromolecules synthetic binders and catalysts have been generated using "evolution in the test tube" of nucleotides (aptamers) or the total synthesis of (molecularly imprinted) polymers (MIPs).

Molecularly Imprinted Polymers To mimic the active sites of proteins synthetic polymers have been "imprinted" by the analyte during the formation of a polymeric network. Before polymerization the target interacts by covalent and non-covalent binding with the chemically active moieties of the "functional monomers." This arrangement is fixed in the subsequent polymerization of functional monomers or the reaction with a crosslinker. After polymerization the template molecules are removed, providing binding sites that are complementary in size and shape to the template and, thus, the template preferentially rebinds to these cavities.

Whilst nature has a spectrum of 20 amino acids as building elements of proteins, in molecularly imprinted polymers (MIPs) only one or up to three different functional monomers have been used. The most common functional monomers applied in thermal or photo-polymerization are methacrylic acid, vinylimidazole, vinylpyridine, and their derivatives. Electropolymerization of pyrrole, scopoletin, phenylenediamine, thiophene, *p*-aminophenyl-boronic acid, and their derivatives in the presence of the target molecule allows the preparation of MIPs without any crosslinker.

Molecularly imprinted polymers have been successfully developed for low molecular weight substances. The preparation of MIPs for biomacromolecules like proteins is, however, still a challenge. In contrast to the synthesis of MIPs for low molecular weight molecules all steps, including polymerization, should be realized in aqueous media. For effective removal of the target protein from the MIP and accessibility for rebinding, the binding sites should be located on the

surface of the polymers. For this reason, surface imprinting techniques have been developed, where the template is imprinted in polymer films that are thinner than the protein molecule.

MIPs have found routine application in affinity chromatography and it has been expected that they will replace antibodies in bioanalysis. They are especially appropriate for toxic targets and for measurements in organic solvents, that is, in extracts of environmental samples or foods. Despite major efforts, the performance of MIPs is still behind that of antibodies.

In addition to binding MIPs, catalytically active MIPs have been developed for application in sensors and syntheses. In these MIPs either an analogue of the transition state of the catalysed reaction is used as the template or metal complexes are combined with binding sites for the substrate. So far, however, the catalytic activity of MIPs has in general been between 1% and 10% of their biological counterparts. On the other hand, MIPs have been prepared as catalysts for reactions where no enzymes are available.

Molecularly imprinted polymers have the potential to replace in future biological recognition elements, especially in affinity chromatography but also in biosensors and biochips for the measurement of low molecular weight substances, proteins, viruses, and living cells.

Aptamers Aptamers are single stranded DNA or RNA molecules that are prepared by *in vitro* selection from libraries of synthetic oligonucleotides. This combinatorial approach is known as SELEX (systematic evolution of ligands by exponential enrichment). It consists of the following steps:

1. The generation of DNA libraries by automated DNA synthesis. The diversity of the individual oligonucleotides is typically between 10^{13} and 10^{15}, which is even higher than in our immune system. The randomized region consists of 30–60 nucleotides that are flanked on both sides by primer sequences for amplification by the polymerase chain reaction (PCR).
2. Selection of the analyte–binding oligonucleotides, which consists of the interaction of the total library with the immobilized target analyte and the removal of non-binding species by elution.
3. Isolation of the bound aptamers under stringent conditions and amplification by PCR.

Within 6–10 cycles with increasing stringency of the binding conditions aptamers with affinities comparable to that of antibodies can be isolated. Aptamers even distinguish the chirality of molecules or the secondary structure of proteins. The exchange rate of the analyte–aptamer complex can be considerably higher than for antibodies. This behaviour can allow for sensor regeneration and even for "online measurements."

Once the sequence of an aptamer is identified, it can be synthesized with high purity without consumption of the target. Aptamers can be easily modified by tags or markers including electrochemical indicators and fluorescence probes. Furthermore, aptamers can be easily regenerated without loss of affinity and selectivity.

On the other hand, immobilization of the aptamer to the transducer surface or alteration of the environment can change the structure and therefore influence the interaction with the target. Problems can arise from degradation by the ubiquitous DNases.

Today, aptamers for almost 300 different analytes, including metal ions, organic dyes, drugs, amino acids, cofactors, antibiotics, proteins, and nucleic acids have been synthesized. They are still applied in lateral flow devices and different sensor configurations.

17.4 From Glucose Enzyme Electrodes to Electronic DNA Biochips

Biosensors have developed in the following directions:

- diversification of biorecognition elements and signal transducers,
- increased integration,
- miniaturization,
- as arrays.

The biochemical recognition of the analyte is still the key part of the procedure. The spatial integration with signal transduction led to biosensors and biochips. In addition, development of the transducer has influenced mainly the evolution of rapid biosensors. The oxygen electrode of Leland Clark Jr. and pH sensors initiated the final successful biosensor development.

After the electrochemical start with enzyme electrodes, in analogy to classical photometry, the biochemical recognition was optically monitored: optical fibers covered with pH indicators have been used to show enzyme reactions (e.g., the urease-catalyzed splitting of urea). In addition, the fluorescent monitoring of NADH conversion has also followed this line. Breakthroughs, however, have been made only after stable fluorescent dyes became available. This has happened for immunosensors and over a broad range for DNA chips. Fluorescent-labeled binding partners are now dominant in DNA sensors and nucleic acid assays.

With the advent of Biacore on an SPR basis by Pharmacia, label free assays have gained importance. This holds for optical methods (like reflectometric interference spectroscopy), but also for piezoelectric quartz crystals (PQC).

Surface Plasmon Resonance (SPR), Sections 5.3.3, 16.6, 32.5.5

A unique ability of biosensors is the online concentration measurement of key metabolites by implanted sensors. This allows the time-resolved indication of essential metabolites (e.g., glucose, lactate, and glutamate), of drugs, and of hormones. Especially important is the measurement of short-lived mediators like superoxide or nitric oxide in body fluids or on the surface of cells.

After the breakthrough of fluorescence-based DNA chips for sequence analysis, the development of new technologies for sequencing single nucleic acid molecules brought the *1000 dollar genome* close to reality: In the approach of Pacific Biosciences the liberation of fluorescent side products during the synthesis of double stranded DNA by single DNA polymerase molecules is indicated by a zero-mode wave guide. Oxford Nanopore's system uses nanopore sequencing to rapid read DNA sequences. A strand of DNA is fed through a biological pore and the bases are identified by measuring the electrical resistance as they pass through the pore.

In these new analytical tools the recognition element – either a nanopore or the DNA polymerase – is spatially integrated with the transducer and, thus, they are biosensors according to the definition of IUPAC.

17.5 Resume: Biosensor or not Biosensor is no Longer the Question

55 years after the first report about the enzyme electrode by L.C. Clark and almost 20 years after the biosensor definition by IUPAC the criteria for this analytical tool are no longer unequivocal.

Protein engineering and biomimetic approaches brought about new recognition elements which are complementary to the traditional biological specifiers like enzymes, antibodies, receptors and nucleic acids. In this way, the fields of chemosensors and biosensors are merging.

Artificial binder molecules which are called as aptamers have been developed by the synthesis of large combinatorial libraries of nucleotides or peptides in combination with the selection for optimal affinity and specificity. A comparably large spectrum of fully synthetic Molecularly Imprinted Polymers has been described in literature as mimics of antibodies (plastibodies) or enzymes. In spite of great efforts and impressive new functional principles -esp. in combination with nanotechnologies- the analytical application is still limited to a few examples.

The impressive achievements of silicon technology, paper-based microfluidics and 3D-printing abolished the distinctions between sensor configurations and microanalytical systems. The most important criterion for the biosensor- spatial integration of recognition element and transducer-holds also for the new analytical microsystems.

Further Reading

Gauglitz, G. (2010) Direct optical detection in bioanalysis: an update. *Anal. Bioanal. Chem.*, **398**, 2363–2372.

Hall, E.A.H. (1991) *Biosensors*, Prentice Hall, New Jersey.

Heller, A. and Feldman, B. (2008) Electrochemical glucose sensors and their application in diabetes management. *Chem. Rev.*, **108**, 2482–2505.

Hussain, M., Wackerlig, J., and Lieberzeit, P.A. (2013) Biomimetic strategies for sensing biological species. *Biosensors*, **3**, 89–107.
Li, S., Cao, S., Whitcombe, M.J., and Piletsky, S.A. (2014) Size matters: challenges in imprinting macromolecules. *Prog. Polym. Sci.*, **39**, 145–163.
Menger, M., Yarman, A., Erdőssy, J., Yildiz, H.B., Gyurcsányi, R.E., and Scheller, F.W. (2016) MIPs and aptamers for recognition of proteins in biomimetic sensing. *Biosensors*, **6**, 35.
Renneberg, R. (2009) *Bioanalytik für Einsteiger*, Spektrum Akademischer Verlag, Heidelberg.
Renneberg, R. and Lisdat, F. (eds) (2008) *Biosensing for the 21st Century*, Advances in Biochemical Engineering/Biotechnology, vol. 109, Springer-Verlag, Berlin, Heidelberg.
Scheller, F.W., Yarman, A., Bachmann, T., Hirsch, T., Kubick, S., Renneberg, R., Schumacher, S., Wollenberger, U., Teller, C., and Bier, F.F. (2013) Future of biosensors: a personal view, in *Biosensors Based on Aptamers and Enzymes*, Springer, Berlin, Heidelberg.
Schalkhammer, Th.G.M. (2002) *Analytical Biotechnology*, Birkhäuser-Verlag, Basel.
Turner, A. (2013) Biosensors: then and now. *Trends Biotechnol.*, **31**, 119–120.
Willner, I. and Katz, E. (2005) *Bioelectronics*, Wiley-VCH Verlag GmbH, Weinheim.
Wollenberger, U., Renneberg, R., Bier, F.F., and Scheller, F.W. (2003) *Analytische Biochemie – Eine praktische Einführung in das Messen mit Biomolekülen*, Wiley-VCH Verlag GmbH, Weinheim.

Part II

3D Structure Determination

Magnetic Resonance Spectroscopy of Biomolecules

18

18.1 NMR Spectroscopy of Biomolecules

Tad A. Holak[1], Martin Schwalbe[2], and Markus Zweckstetter[3]

[1]Jagiellonian University, Faculty of Chemistry, Ingardena 3, 30-060 Cracow, Poland
[2]Institute of Experimental Internal Medicine, Medical Faculty, Otto von Guericke University, Leipziger Straße 44, 39120 Magdeburg, Germany
[3]German Center for Neurodegenerative Diseases (DZNE), & Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany

The phenomenon of nuclear magnetic resonance (NMR) was discovered in 1945 and describes the absorption of electromagnetic waves by nuclei in a magnetic field. A homogeneous magnetic field causes the energy levels of a nucleus to split into several states. Irradiation of radio waves induces a transition between these states if the frequency of the radio waves matches the energy difference between the levels. Compared with other common optical spectroscopic methods NMR spectroscopy differs in two major points: (i) the energy levels only split in the presence of a magnetic field and (ii) the nuclear spins interact with the magnetic but not the electric part of the electromagnetic radiation.

With the introduction of advanced techniques like pulsed Fourier-transform NMR and multidimensional NMR spectroscopy in the late 1960s, the importance of NMR has increased enormously. Today, NMR spectroscopy is mainly used in organic chemistry and biochemistry for structural analysis. It is the only method capable of determining the structure of biomolecules (proteins or nucleic acids) on an atomic level in solution. Especially for structural characterization of the so-called intrinsically disordered proteins (IDPs) as well as membrane proteins, NMR spectroscopy is indispensable (Section 18.1.7, Intrinsically Disordered Proteins). In addition to structure determination, various time-dependent phenomena are investigated by NMR spectroscopy, including reaction kinetics, intramolecular dynamics of proteins and nucleic acids, molecular recognition processes, and the folding and misfolding of proteins. NMR spectroscopy can also be applied to solid-phase samples such as tissue and protein aggregates. This branch of NMR spectroscopy is called solid-state NMR. NMR spectroscopy is thus an extremely versatile technique in biochemistry and biophysics.

Limitations of NMR spectroscopy result from the relatively low sensitivity of the technique as well as from the high degree of complexity and information content in the NMR spectra. However, advancements in electronic components of the spectrometer and increasing magnetic field strengths have steadily improved the sensitivity and resolution. New NMR experiments extend the range of applications (to larger proteins) and allow the determination of additional structural and dynamic parameters, for example, direct observation of hydrogen bonds and the detection of internal motions from picoseconds to milliseconds and hours. Recombinant protein expression techniques enable spectroscopists to produce the required protein amounts and to incorporate the rare isotopes ^{13}C and ^{15}N into the protein. The detection of these nuclei simplifies the NMR spectra and provides additional parameters that give insight into the

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

Table 18.1 Nuclear spin, natural abundance, gyromagnetic ratio γ, and relative and absolute receptivity of some important nuclei for NMR spectroscopy of biologic macromolecules.
Adapted from Friebolin, H. (2011) *Basic One- and Two-dimensional NMR Spectroscopy*, VCH-Verlagsgesellschaft, Weinheim.

Nucleus-isotope	Spin *I*	Natural abundance (%)	Gyromagnetic ratio γ [a] (10^7 rad T^{-1} s^{-1})	Relative receptivity[b]	Absolute receptivity[c]
^{1}H	$^{1}/_{2}$	99.980	26.7519	1.00	1.00
^{2}H	1	0.016	4.1066	9.65×10^{-3}	1.54×10^{-6}
^{12}C	0	98.900	—	—	—
^{13}C	$^{1}/_{2}$	1.108	6.7283	1.59×10^{-2}	1.76×10^{-4}
^{14}N	1	99.630	1.9338	1.01×10^{-3}	1.01×10^{-3}
^{15}N	$^{1}/_{2}$	0.370	–2.712	1.04×10^{-3}	3.85×10^{-6}
^{16}O	0	99.960	—	—	—
^{17}O	$^{5}/_{2}$	0.037	–3.6279	2.91×10^{-2}	1.08×10^{-5}
^{31}P	$^{1}/_{2}$	100.000	10.841	6.65×10^{-2}	6.65×10^{-2}

a) The γ-values are from Harris, R.K. (1983) Nuclear Magnetic Resonance Spectroscopy. A Physico-chemical View. Pitman, London.
b) Detection efficiency at constant magnetic field and equal number of nuclei.
c) Product of relative sensitivity and natural abundance.

structure and dynamics of proteins. Nowadays NMR spectroscopists routinely study proteins up to a molecular weight of 30 kDa. Dedicated techniques allow even the analysis of MDa protein complexes (Section 18.1.7, Structure and Dynamics of High Molecular Weight Systems and Membrane Proteins).

The proton is the most important nucleus for NMR spectroscopy because it naturally occurs in high abundance and is also the most sensitive nucleus. However, other magnetically active nuclei exist (Table 18.1), of which ^{13}C, ^{15}N, and ^{31}P are important for biomolecular NMR.

18.1.1 Theory of NMR Spectroscopy

The nuclear spin is a quantum mechanical property and cannot be explained or adequately described within the framework of classical theory. The formalism necessary for this would, however, be beyond the scope of this chapter. Nevertheless, we can illustrate many aspects by a classical approach if we introduce quantum mechanical properties, such as scalar coupling between nuclear spins, *ad hoc* through simple rules. After a short introduction to the fundamental properties of nuclear spins, we will restrict our description to a macroscopic magnetization comparable to that of a tiny bar magnet.

Nuclear Spin and Energy Quantization The phenomenon of nuclear magnetic resonance relies on the interaction of the magnetic moment μ of a nucleus with the external magnetic field. The cause of the magnetic moment is the quantum mechanical property of the nucleus that is called spin. The nuclear spin results from the combination of the individual spins of the protons and neutrons (each with spin ½) that build the nucleus. Pairs of protons or neutrons tend to combine their spins in an antiparallel configuration, thus canceling most of their contributions. For example, due to the even number of protons and neutrons (six each), the twelve spin ½ contributions of the carbon isotope ^{12}C add up to a net nuclear spin of zero. In contrast, the nucleus of hydrogen consists of only one proton and therefore has a nuclear spin of ½.

The name "spin" might suggest that it could be an intrinsic angular momentum of the particle arising from its rotation because it is well known that the rotation of a charged particle results in a magnetic moment. Equation 18.2 below appears to support this relationship between nuclear spin and magnetic moment. However, the quantum mechanical spin has absolutely nothing to

do with spatial self-rotation and is an intrinsic property of the particle. Nevertheless, a commonality between the nuclear spin and classical self-rotation exists that justifies the name "spin." Mathematically, the quantum mechanical spin (of protons, electrons, etc.) and the classical angular momentum show a similar behavior, following angular momentum commutator relations from which many important properties can be derived. From Equation 18.1 it is evident that the nuclear spin is quantized:

$$|\boldsymbol{j}| = \sqrt{I(I+1)}\hbar \tag{18.1}$$

where $|\vec{J}|$ is the spin angular momentum, I the nuclear spin quantum number (that can adopt values from $I = 0, ½, 1, 1½, \ldots, 6$), and $\hbar = h/2\pi$ is Planck's quantum of action. The spin angular momentum is directly proportional to the magnetic moment μ:

$$\boldsymbol{\mu} = \gamma \boldsymbol{j} \Rightarrow |\boldsymbol{\mu}| = \gamma\hbar\sqrt{I(I+1)} \tag{18.2}$$

The constant of proportionality γ (denoted as the gyromagnetic ratio) is a characteristic constant for each type of nucleus. The sensitivity in NMR mainly depends on γ – a large gyromagnetic ratio means a high sensitivity of the respective isotope (Table 18.1).

By convention, in an external magnetic field the component of J along the direction of the field J_z (if the magnetic field is along the z-axis) is an integer or semi-integer multiple of Planck's quantum of action:

$$J_z = m\hbar \Rightarrow \mu_z = m\gamma\hbar \tag{18.3}$$

in which the magnetic quantum number m can adopt values from $-I$ to $+I$. An external magnetic field causes a splitting of the energy levels of the nucleus. For example, considering a hydrogen nucleus with a spin ½, two possible states result according to a parallel (↑, $m = ½$) and an antiparallel (↓, $m = -½$) orientation of the nuclear spin relative to the external field (Figure 18.1). The energy E of these levels is given by the classic formula for a magnetic dipole in a homogeneous magnetic field of strength B_0:

$$E = -\mu_z B_0 = -m\gamma\hbar B_0 \tag{18.4}$$

The magnetic moment of each nucleus rotates in a precessing motion around the direction of the field B_0 with a frequency denoted as the Larmor frequency ω_0 or ν_0. This frequency corresponds to the resonance frequency of the nucleus and thus to the transition frequency between the energy levels:

$$\gamma\hbar B_0 = \Delta E = h\nu = \hbar\omega_0 \Rightarrow \omega_0 = \gamma B_0 \quad \text{with} \quad \omega_0 = 2\pi\nu_0 \tag{18.5}$$

Therefore, the Larmor frequency depends on the gyromagnetic ratio γ and the strength of the magnetic field (Figure 18.2). For example, at a magnetic field strength of 18.8 T (approx. 380 000 times stronger than the earth magnetic field) the Larmor frequency of protons is 800 MHz (5.027 Mrad s^{-1}), which corresponds to the frequency range of radio waves.

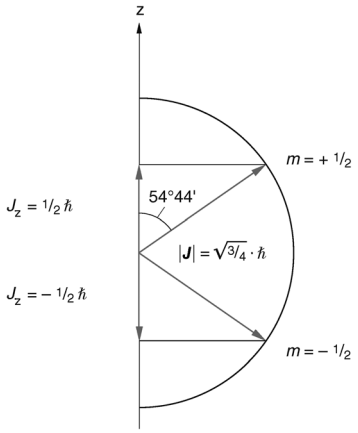


Figure 18.1 Quantization of the direction of the nuclear spin angular momentum J in a homogeneous magnetic field for nuclei with $I = ½$. The magnetic field is along the z-axis. Because $J_z = 1/2\hbar$ and $|\boldsymbol{j}| = \left(\sqrt{3}/2\right)\hbar$, $\boldsymbol{J}$ adopts an angle of 54°44′ (= arccos(Jz/|J|) with respect to the z-axis.

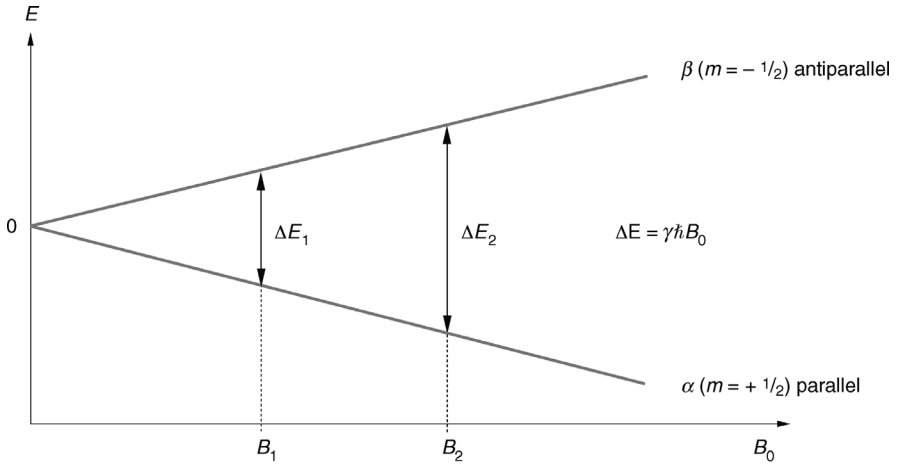


Figure 18.2 The energy difference ΔE between the two energy levels of a nuclear spin ½ depends linearly on the magnetic field strength B_0.

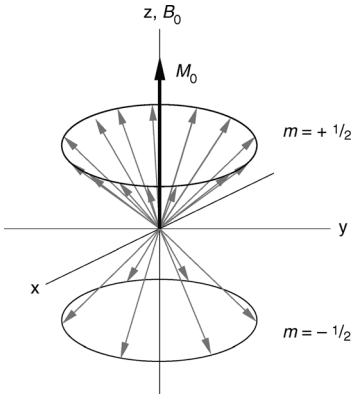


Figure 18.3 Illustration of the macroscopic *z*-magnetization at thermal equilibrium. The combined *z*-components of the magnetic moments of individual spins result in the macroscopic magnetization vector M_0 (bold arrow). Because $N\uparrow > N\downarrow$, the vector points along the +*z*-axis. The *x*- and *y*-components are uniformly distributed on the surface of the double precession cone and do not produce a macroscopic magnetization.

Populations and Equilibrium Magnetization The common concentration of samples analyzed by NMR ranges from submicromolar to moles per liter (for proteins 0.01–10 mM). The nuclear spins of these molecules align themselves independent of each other either parallel or antiparallel along the external magnetic field. For nuclei with a spin ½ the population ratio of the moments (the number of nuclei $N_\downarrow$ that align antiparallel to the main field over the number with parallel orientation $N_\uparrow$) obeys the Boltzmann distribution:

$$\frac{N_\downarrow}{N_\uparrow} = \exp\left(\frac{\Delta E}{kT}\right) = \exp\left(\frac{\gamma\hbar B_0}{kT}\right) \tag{18.6}$$

where k is the Boltzmann constant and T is the absolute temperature. Because the energy difference between both levels is orders of magnitude smaller than the thermal energy (kT), both levels are almost equally occupied. For example, for protons at 300 K and a magnetic field of 18.8 T (800 MHz) the excess population in the lower energy level amounts to only 1.3 in 10 000 particles (i.e., $N_\downarrow = 0.99987 \times N_\uparrow$). This is the main reason why NMR spectroscopy is so insensitive compared with other spectroscopic methods.

Even this tiny difference suffices to give rise to a macroscopic bulk magnetization M_0, which results from the combined magnetic moments of the individual nuclear spins. The macroscopic magnetization of spin ½ particles is then given by:

$$M_0 = N\frac{\gamma^2\hbar^2 B_0 I(I+1)}{3kT} = N\frac{\gamma^2\hbar^2 B_0}{4kT}, \quad \text{for} \quad I = \frac{1}{2} \tag{18.7}$$

This shows that the magnitude of the equilibrium magnetization depends on the magnetic field strength B_0, the number of spins N, and the temperature T of the sample. Varying each of these parameters can enhance the observable signal (which is the reason for the ongoing development of magnets with increasing field strengths). Importantly, the spectrometer records the time development of exactly this macroscopic bulk magnetization. Moreover, the classical theory of NMR spectroscopy normally considers this macroscopic magnetization as due to its more descriptive behavior than to that of individual spins.

In equilibrium, magnetization M_0 exists only along the axis of the main field (by convention the *z*-direction, i.e., $M_z = M_0$). The transverse *x*- and *y*-components of the magnetic moments are uniformly distributed and add up to zero ($M_x = M_y = 0$) (Figure 18.3).

The Bloch Equations According to the Bloch equation, the change of the magnetization vector $\boldsymbol{M}$ over time results from the interaction of the magnetization with an effective magnetic field $\boldsymbol{B}_{\text{eff}}$:

$$\frac{\mathrm{d}\boldsymbol{M}}{\mathrm{d}t} = \gamma(\boldsymbol{M} \times \boldsymbol{B}_{\text{eff}}), \quad \boldsymbol{B}_{\text{eff}} = \underbrace{\left(\boldsymbol{B}_0 + \frac{\omega_0}{\gamma}\right)}_{0} + \boldsymbol{B}_1 = \boldsymbol{B}_1 \tag{18.8}$$

To simplify the mathematical treatment of the magnetization vector, we will introduce a rotating coordinate system, which precesses about the *z*-axis with the Larmor frequency of the nuclei ($\omega_0 = -\gamma B_0$). In this coordinate system all nuclear spins rotating with the Larmor frequency appear stationary. This concept should be familiar for us because we all live in a rotating coordinate system — the Earth. A person standing on the equator moves for an observer in space with a speed of approx. 1700 km h^{-1}. Consider that this person on earth throws a ball straight up into the air and then sees it falling down again. For him or her the ball moves on a simple straight vertical path; however, for our observer in space this ball would move on a complicated trajectory.

In the rotating coordinate system the contribution of the main magnetic field $\boldsymbol{B}_0$ to $\boldsymbol{B}_{\text{eff}}$ cancels for nuclei with Larmor frequency $\boldsymbol{\omega}_0$. Moreover, if only the main magnetic field $\boldsymbol{B}_0$ is applied, then also $\boldsymbol{B}_{\text{eff}}$ is zero and the magnetization vector becomes time-independent. However, if an additional field $\boldsymbol{B}_1$ is applied perpendicular to the main magnetic field $\boldsymbol{B}_0$, then $\boldsymbol{B}_{\text{eff}} = \boldsymbol{B}_1$. The magnetization vector will now precess around the axis of the $\boldsymbol{B}_1$ field if its frequency matches the Larmor frequency of the nuclei ($\omega_{\text{rf}} = \omega_0$, the resonance condition). Thus, the $\boldsymbol{B}_1$ field induces a rotation of the magnetization vector $\boldsymbol{M}_z$ from the equilibrium position to the transverse plane (cross product). Physically, the $\boldsymbol{B}_1$ field is nothing other than a

short radio frequency pulse. For example, a so-called excitation pulse or 90°-pulse completely converts *z*- into *y*-magnetization if the $\boldsymbol{B}_1$ field along the *x*-axis is of appropriate strength and duration (Figure 18.4).

How can we imagine the origin of the *y*-magnetization for individual spins after a 90°-pulse? (i) Both energy levels are equally populated because $M_z = 0$. (ii) The magnetization dipoles of individual spins are not uniformly distributed around the *z*-axis, but small parts precess bundled about the *z*-axis. It is this state of phase coherence that gives rise to the macroscopic *y*-magnetization (Figure 18.5).

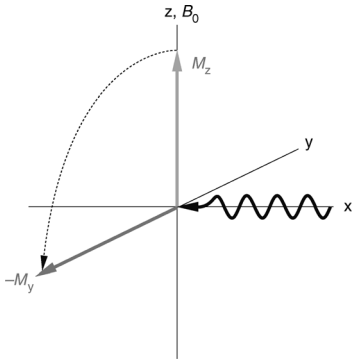


Figure 18.4 Effect of a 90°-pulse on *z*-magnetization. A pulse about the *x*-axis (bold wavy arrow) rotates the equilibrium *z*-magnetization (grey arrow) by 90° counterclockwise around the *x*-axis and creates *–y*-magnetization (grey arrow).

Relaxation The Bloch equation shown above is incomplete because it predicts an infinitely long precession of the magnetization vector once the sample has been excited. However, the transverse magnetization after the 90°-pulse corresponds to a non-equilibrium state from which the system will return to its thermodynamic equilibrium after a short while. Therefore, Bloch defined two different relaxation time constants denoted as longitudinal (T_1) and transverse (T_2) relaxation time constants. Assuming that the respective relaxation processes follow first-order rate equations, the Bloch equations are modified to:

$$\frac{dM_{x,y}}{dt} = \gamma(M \times B)_{x,y} - \frac{M_{x,y}}{T_2} \tag{18.9}$$

$$\frac{dM_z}{dt} = \gamma(M \times B)_z + \frac{M_0 - M_z}{T_1} \tag{18.10}$$

The equations indicate that due to relaxation the transverse components (M_x, M_y) decay to their equilibrium value of zero, whereas longitudinal magnetization M_z approaches its equilibrium value M_0. In high resolution NMR spectroscopy, the T_1 relaxation time constants for protons are in the range of one to several seconds. Usually, T_2 is similar to T_1 (small molecules); however, for large molecules like proteins T_2 is much smaller than T_1. From this fact results the well-known size limitation of NMR spectroscopy to proteins, because the enhanced T_2 relaxation in large proteins decreases the resolution and sensitivity of their NMR spectra (Section 18.1.2, Spectral Parameters). Notably, relaxation effects can often be ignored during the radiofrequency pulses, because relaxation times are long compared to commonly used pulse lengths (10–50 µs).

Relaxation is caused by different time-dependent interactions (e.g., dipolar couplings) between the spins and their environment (T_1) and between the spins themselves (T_2). Historically, T_1 is also referred to as the spin–lattice and T_2 as the spin–spin relaxation time constant. The relaxation time constants depend on several factors including the Larmor frequency (or magnetic field strength) and the molecular mobility of the molecule in solution. The latter is characterized by the rotational correlation time, $\boldsymbol{\tau_c}$. We will discuss the measurement of relaxation times in more detail later during the analysis of protein dynamics (Section 18.1.7, Determination of Protein Dynamics).

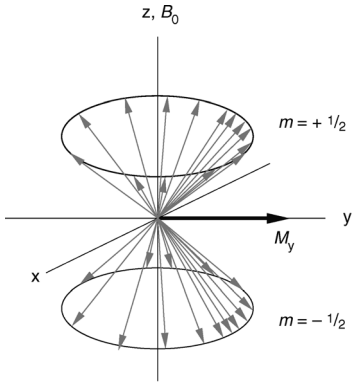


Figure 18.5 Illustration of transverse magnetization. The identical number of nuclear spins (grey arrows) in both energy levels shows that they are equally populated. Some nuclear spins precess bundled (or in phase) about the direction of the B_0 field. Their magnetic moments add up to the macroscopic M_y-magnetization (bold arrow).

Pulsed Fourier Transformation Spectroscopy Modern NMR spectrometers operate with a technique called pulsed Fourier Transformation NMR (FT-NMR) spectroscopy. This technique replaced older NMR methods (e.g., continuous wave NMR spectroscopy) because it greatly improved the sensitivity and resolution, and also facilitated the development of multidimensional NMR methods (Sections 18.1.3 and 18.1.4). In pulsed FT-NMR spectroscopy all nuclei are excited *simultaneously* through a radio frequency pulse. The radio transmitter works at a fixed frequency ν_0 and would therefore excite only nuclei with this Larmor frequency (resonance condition!). However, the pulse duration is inversely proportional to the frequency bandwidth (and thus the energy of the radiation). If therefore a very short pulse (a few microseconds) is emitted, then the pulse is less "frequency selective." It contains a broad excitation band around ν_0 and excites the Larmor frequencies of all nuclear spins in the sample at once.

Strictly speaking, the flip angle through which the pulse rotates the bulk magnetization depends on the *offset* (or distance) of the Larmor frequencies from the transmitter frequency. For nuclei that are *off-resonance*, the effective field $\boldsymbol{B}_{\text{eff}}$ is not collinear with the $\boldsymbol{B}_1$ field as in the resonant case. As a result, the flip angle for *off-resonance* nuclei decreases with increasing *offset*. However, the projection of the transverse magnetization on the *y*-axis depends on the sine

of the flip angle. For example, even for an 80° flip angle for far *off-resonance* nuclei the projection is 98.5% of that of a 90°-pulse – a more than acceptable result for NMR spectroscopy.

After this excitation pulse, the different nuclei precess with their different Larmor frequencies about the z-axis. According to Maxwell's equations, a rotating magnetic moment creates a changing magnetic field that induces a current in a wire coil. In the NMR spectrometer, a sensitive receiver coil records this small oscillating current. Through T_2 relaxation the induced current decays over time, which is why the recorded data is called free induction decay (FID). Because the current is detected in a time-dependent manner (and not frequency–selective), the acquired signal is a superposition of all frequencies that have been excited with the pulse. The mathematical operation Fourier transformation converts this time-domain signal into the frequency-domain or spectrum.

In analogy to other spectroscopic methods one could imagine the resonance phenomenon differently. In the resonant case nuclei absorb the radio radiation if the frequency of the radio pulse matches their Larmor frequency. After the pulse all excited nuclear spins simultaneously emit the absorbed radio radiation, which is then detected. Therefore, the pulsed Fourier transformation method is often compared to the tuning of a bell. In principle, one could determine the individual tones, which make up the sound of the bell, in the fashion of a "continuous wave-experiment." The bell is sequentially excited with all sonic frequencies through a loudspeaker from the lowest tones to the limit of ultrasound, and the reaction of the bell is measured with a microphone. This procedure is extremely cumbersome and every bell founder knows that it can be done quicker; one simply takes a hammer and strikes! The sound of the bell contains all tones at once and every human being can analyze the sound with his or her ears (an ingeniously "constructed" biological tool for Fourier transformation). Note, however, that no frequency-dependent detection occurs on modern NMR spectrometers.

18.1.2 One-Dimensional NMR Spectroscopy

The 1D Experiment With these basic theoretical principles we are able to understand the simplest variant of NMR spectroscopy – the one-dimensional (1D) experiment (Figure 18.6). Each 1D NMR experiment consists of two parts: *preparation* and *detection*. During the preparation the spin system is brought to a defined state; during the detection the response to the preparation is recorded.

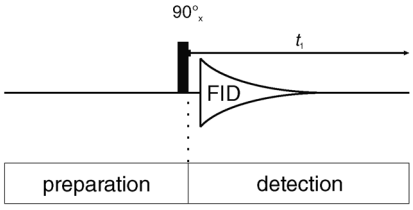


Figure 18.6 Schematic illustration of the pulse sequence for a one-dimensional NMR experiment. A 1D experiment consists of two parts, the preparation and the detection. In the simplest case, the preparation consists of a single 90° pulse (black bar). Subsequently, the response of the spin system (FID) to this pulse is recorded during the detection period.

Preparation of the spin system consists in the simplest case of a short, hard excitation pulse (ca. 10 µs) that creates transverse magnetization (compare Figure 18.4). The resulting FID is recorded and saved during the detection period. After a short waiting time (the relaxation delay) that allows the magnetization to return to its equilibrium value through T_1 relaxation, the experiment can be repeated multiple times. The individual data of the measurement are then added together, increasing the signal-to-noise ratio. Multiplication of the FID data with window functions can enhance either the sensitivity or the resolution of the spectrum. Additionally, this operation suppresses artifacts that arise from the subsequent Fourier transformation, which converts the FID (time-domain) into the NMR spectrum (frequency domain).

Spectral Parameters We will discuss the different NMR spectral parameters (chemical shift, scalar couplings and line width) on the basis of the simple 1D NMR spectrum of ethanol

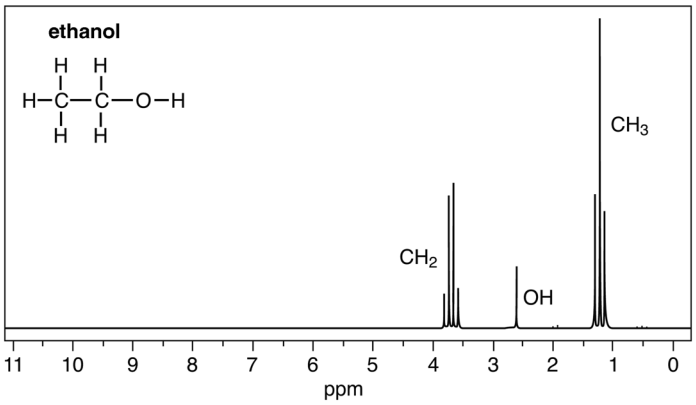


Figure 18.7 One-dimensional ^{1}H NMR spectrum of ethanol. The signal of the methylene group is split into four lines with an intensity ratio of 1 : 3 : 3 : 1 (quartet); that of the methyl group into three lines with a ratio of 3 : 6 : 3. According to the number of hydrogen atoms, one obtains an intensity ratio (= integrals of the signals) of 2 : 3 for the methylene and methyl signals. The hydroxyl proton rapidly exchanges with hydroxyl protons from other ethanol molecules. Therefore, its signal at 2.6 ppm is much broader relative to the other protons. For the same reason, no coupling occurs between the hydroxyl proton and the methylene protons. Hence, neither the hydroxyl signal is split nor does the hydroxyl proton contribute to an additional splitting of the methylene protons.

(Figure 18.7). This spectrum contains three signals (or peaks) originating from the methyl (CH_3) protons, the methylene (CH_2) protons, and the hydroxyl (OH) proton. Because the protons of the methyl group and of the methylene group, respectively, are each equivalent to each other, they each give rise to only one peak. The two peaks appear as so-called multiplets because their signals are split into several lines by scalar coupling. The integral over the respective multiplets yields the number of protons that give rise to these signals. For ethanol one obtains a ratio of the integrals of 3 : 2 : 1 corresponding to the number of protons that contribute to the respective signals.

Chemical Shift In a molecule the electrons surrounding the nucleus create a weak magnetic field and shield the nucleus slightly from the main field. This shielding depends on the specific chemical environment (i.e., the structure of the molecule) and influences the Larmor frequencies of the nuclei. The effect is called the chemical shift and is one of the fundamental parameters in NMR spectroscopy, because it determines the distinct positions of individual signals in the NMR spectrum. The chemical shift δ of a signal in ppm (parts per million) is defined as:

$$\delta = \frac{\omega_{\text{signal}} - \omega_{\text{reference}}}{\omega_{\text{reference}}} \times 10^6 \text{ ppm} \qquad (18.11)$$

The frequencies are given in ppm instead of Hertz because the former unit is independent of the magnetic field strength. The common reference frequency ($\omega_{\text{reference}}$) on which the chemical shift is based is the signal of the methyl groups of tetramethylsilane (TMS). By definition, it has a chemical shift of 0 ppm. For aqueous solutions of proteins and nucleic acids the preferred standards are the methyl signals of 2,2-dimethyl-2-silapentane-5-sulfonic acid (DSS) or trimethylsilylpropanoic acid (TSP). By convention, the chemical shift is plotted on the x-axis of a NMR spectrum from right to left. One often encounters expressions like "a signal appears at high field" (i.e., at low ppm values) or "downfield shift" (i.e., shift towards higher ppm values). These expressions originate from the time when NMR spectra were acquired at constant transmitter frequency through variation of the magnetic field (continuous wave technique).

The position of a peak in the NMR spectrum provides substantial information about the origin of the respective signal. Many chemical or functional groups display specific chemical shifts (Figure 18.8); for example, the chemical shift of the hydroxyl group of ethanol differs from that of the methyl group (Figure 18.7). For proteins, the chemical shift alone suffices to distinguish between the signals from the H^N, H^α, aromatic and aliphatic protons (Figure 18.12 below). The chemical shift further contains information about the secondary and tertiary structure of a protein, which is very valuable in different stages of structure determination (Section 18.1.6, Determination of the Secondary Structure).

Scalar Coupling In the 1D spectrum of ethanol the signals of the methylene and methyl protons (Figure 18.7) appear as multiplets. This line splitting results from the scalar coupling (or indirect

coupling) between the protons, and is mediated through the electrons in the atomic bonds connecting the nuclei. Besides the nuclear Overhauser effect (NOE), scalar coupling is the most important mechanism in multidimensional NMR spectroscopy (Section 18.1.3) by which magnetization is transferred between nuclei.

Line splitting arises due to the different orientations of a spin ½ to the external magnetic field. Each of the two methylene protons can adopt either a parallel or antiparallel orientation, which corresponds to two different magnetic quantum numbers *m*. The protons of the methyl group, which are scalar coupled to the methylene protons, therefore, "experience" four possible combinations (↑↑, ↑↓, ↓↑, and ↓↓). The orientations ↑↑ and ↓↓ marginally enhance or attenuate, respectively, the external magnetic field and thus shift the resonance frequency of the methyl protons (Figure 18.9). This gives rise to two lines lying symmetrically left and right of the actual resonance frequency. The orientations ↑↓ and ↓↑ are equivalent and compensate their respective enhancements or attenuations of the main field, thus, an unshifted line with twice the intensity results. The generated splitting pattern for the methyl group is called a triplet.

(a)

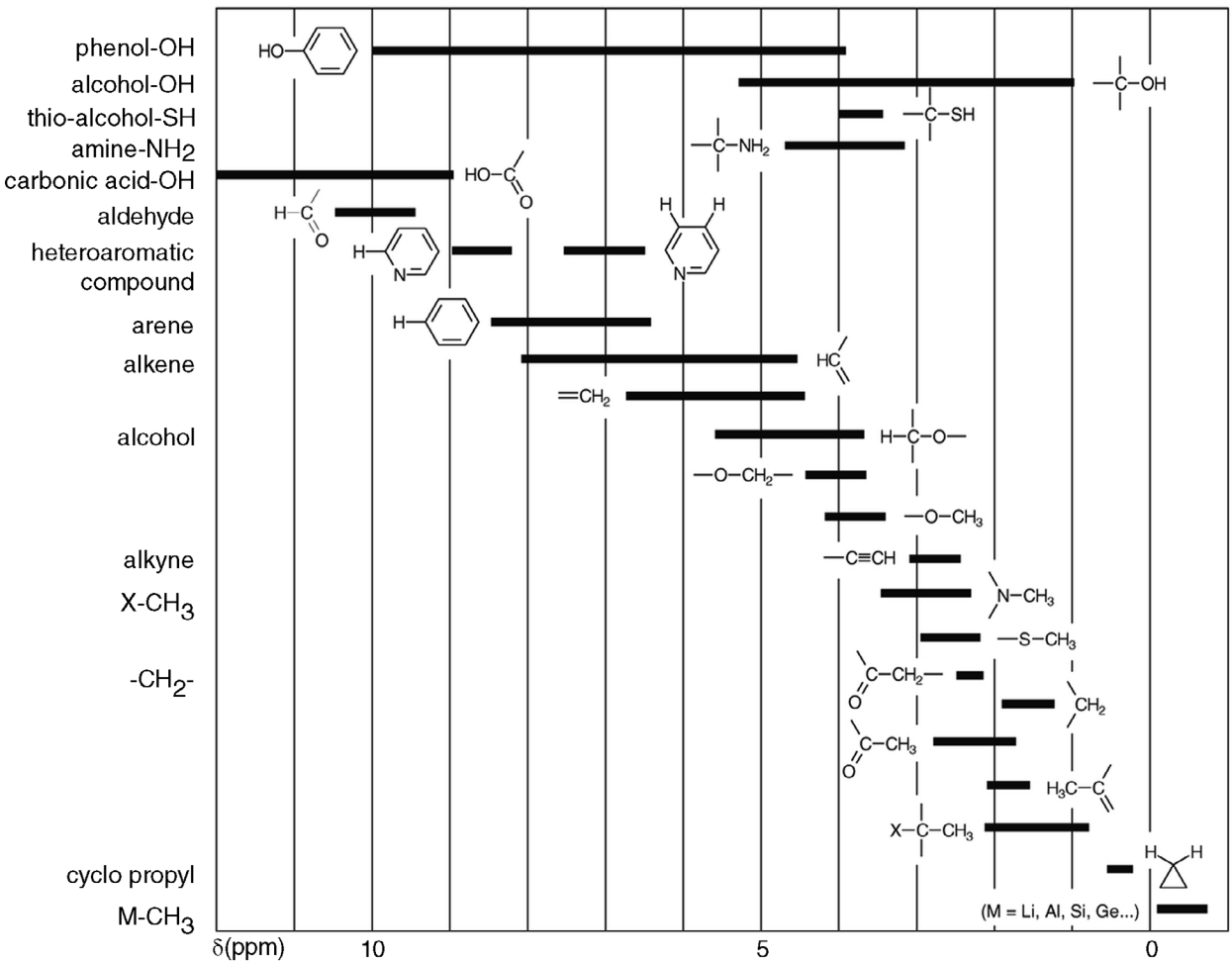


Figure 18.8 (a) Typical proton chemical shift ranges of different chemical groups. Source: adapted from Bruker Almanac, 1993. (b) Typical proton chemical shift ranges of individual amino acids in proteins. Source: adapted from Wishart, D.S., Sykes, B.D., and Richards, F.M. (1991) *J. Mol. Biol.*, **222**, 311–333. (With permission Copyright © 1991 Published by Elsevier Ltd.)

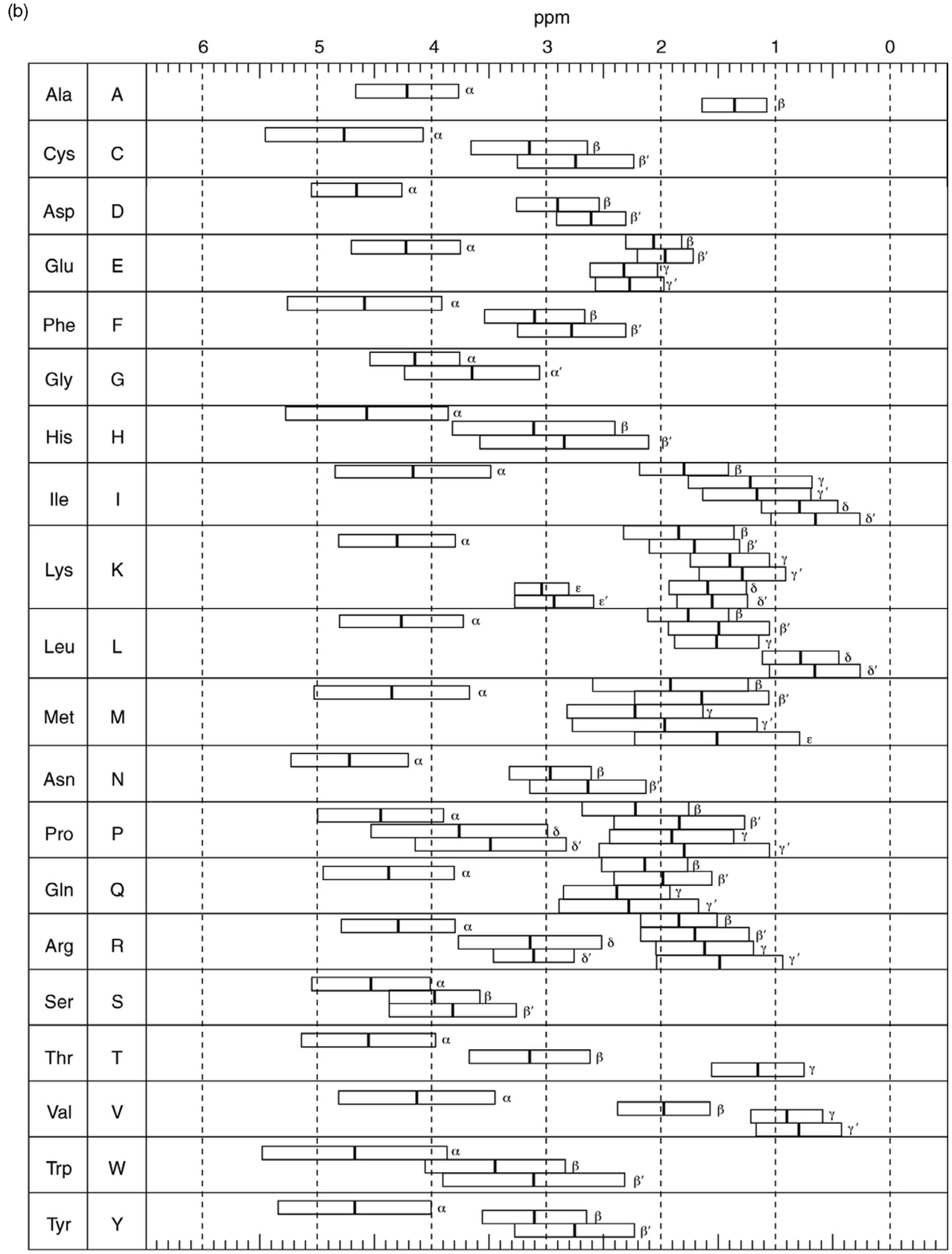


Figure 18.8 (*Continued*)

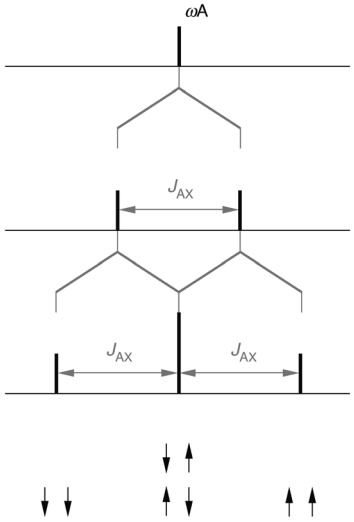


Figure 18.9 Origin of the triplet: In an AX_2 spin system the coupling with two identical nuclei X causes the resonance of nucleus A to split into a triplet. Each of the two X-nuclei can orient itself either parallel or antiparallel to the external magnetic field giving rise to four possible orientations. A parallel orientation enhances, while an antiparallel orientation attenuates, the external magnetic field. Therefore, the line associated with the ↑↑ orientation shifts upfield and the respective line for the ↓↓ orientation shifts downfield. The orientations ↑↓ and ↓↑ are indistinguishable and the respective lines appear at the actual resonance frequency. The single lines of the triplet have an intensity ratio of 1 : 2 : 1.

If two nuclei with spin quantum numbers I and S couple with each other, then the resonance of I splits into $2S+1$ lines and the resonance of S into $2I+1$ lines. If the coupling partners of S consist of several identical I-nuclei, then the resonance S splits into $2nI+1$ lines, where n is the number of identical coupling partners (and vice versa).

The number of individual spin combinations determines the intensity of these lines and follows a binomial distribution that can be illustrated in a Pascal's triangle. For ethanol, two methylene protons (two I-nuclei) couple with three methyl protons (three S-nuclei). Thus, the signal of the methyl group splits in $2\times2\times½+1=3$ lines (triplet), and that of the methylene group in $2\times3\times½+1=4$ lines (quartet). The lines of the triplet have intensity ratios of 1 : 2 : 1; the lines of the quartet of 1 : 3 : 3 : 1. The separation of the lines (in Hz) in a multiplet corresponds to the coupling constant J, which is independent of the applied magnetic field strength. In general, only couplings over one bond (^1J), two bonds (^2J), and three bonds (^3J, or vicinal coupling) are observed (Figure 18.10). An important aspect of vicinal coupling constants is that their magnitude depends on the torsion angle between the two protons. The semiempirical Karplus relationship describes this dependence (Figure 18.11):

$$J(\phi) = A\cos^2(\phi - 60) - B\cos(\phi - 60) + C \qquad (18.12)$$

in which A, B, and C are empirically determined constants that are different for each type of torsion angle (e.g., the ϕ, ψ, and χ angles in proteins). For example, protein structure determination exploits the information about the molecular geometry contained within the $^3J(H^N - H^\alpha)$ coupling constant to restrain the torsion angle ϕ (H^N-N-C^α-H^α).

Line Width The line widths of NMR peaks provide direct evidence about the lifetimes of the respective resonances. The longer the lifetime of a resonance is, the narrower is the line width of its peak (and vice versa). The resonance lifetimes are manly determined by T_2 relaxation and chemical exchange. As mentioned earlier, the short T_2 relaxation time constants of large molecules (proteins > 50 kDa) cause broad line widths and thus result in peaks with low intensity. Similarly, chemical exchange during de- and reprotonation reactions can reduce the lifetime of a proton. For example, the hydroxyl proton of ethanol (Figure 18.7) exchanges with other solvent protons from solvents (in this case other ethanol molecules) and thus possesses a broader line.

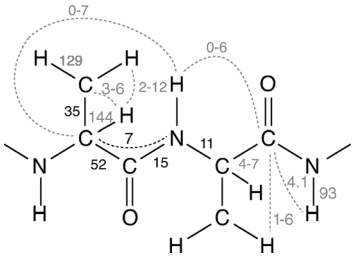
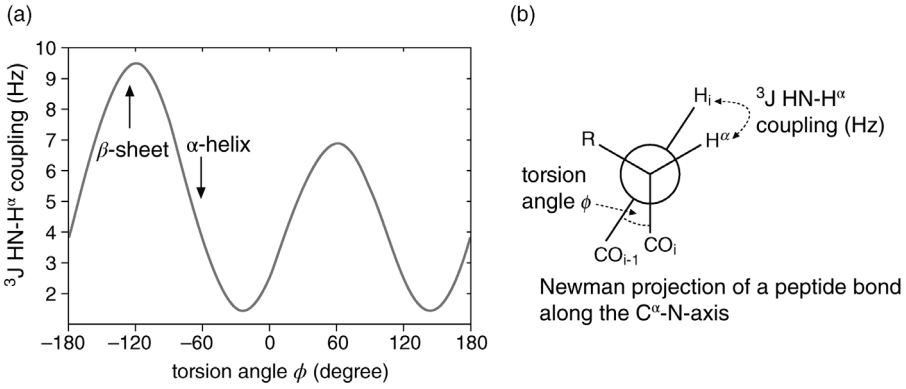


Figure 18.10 Typical values of different coupling constants (in Hz) in the protein backbone. Only coupling constants larger than 5 Hz are considered. Direct CC- and CN-couplings are shown in black; direct CH- and NH-couplings in grey. Black or grey dashed lines highlight indirect CC- and CN-couplings, or indirect CH- and NH-couplings, respectively. Source: adapted from Bystrow, W.F. (1976) *Progr. NMR Spectrosc.*, **10**, 41–81. (With permission Copyright © 1976 Published by Elsevier B.V.)



Karplus relationship: $J(\phi) = 7.97\cos^2(\phi - 60) - 1.26\cos(\phi - 60) + 0.63$
according to Vogeli, B., Ying, J.F., Grishaev, A., Bax, A. Limits on variations in protein backbone dynamics from precise measurements of scalar couplings.(2007) J. Am. Chem. Soc. 129, 9377–9385.

Figure 18.11 (a) Relationship between the $^3J(H^N–H^\alpha)$ coupling constant and the ϕ angle (Karplus relationship). The torsion angle between CO_i and CO_{i-1} (b) is plotted against the coupling constant $^3J(H^N$-$H^\alpha)$. The plot shows that at least two angles correspond to a given coupling constant. The index i in the Newman projection (b) denotes the relative position of the amino acids to each other in the protein sequence.

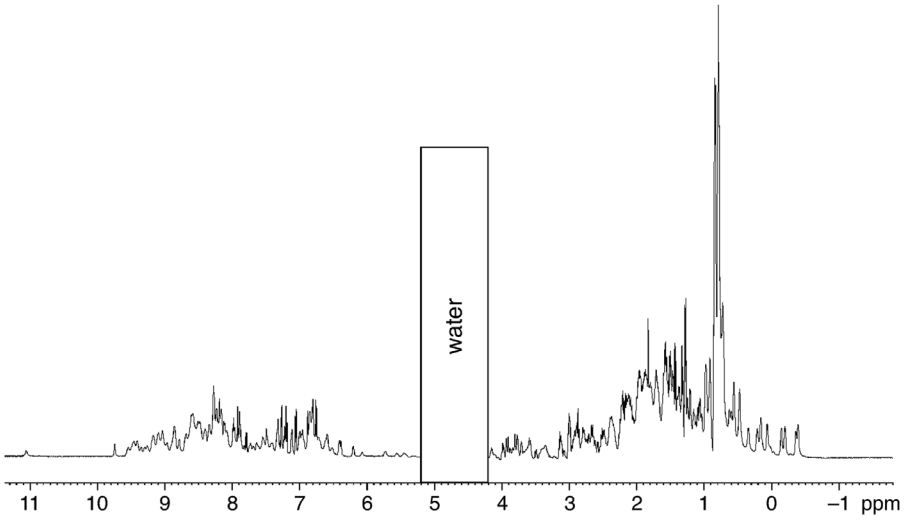


Figure 18.12 1D ^{1}H NMR spectrum of a 14 kDa β-sheet protein with an immunoglobulin fold at 30 °C. The characteristic chemical shift of each proton type facilitates their identification in the different spectral regions. At the left end, at 11 ppm are the peaks of the tryptophan indole protons. The peaks from 10 to 6 ppm are assigned to the amide protons of the protein backbone and of the asparagine and glutamine side chains. Between 7.5 and 6.5 ppm are the aromatic protons, followed by the H^{α} protons from 5.5 to 3.5 ppm. To the right-hand end of the spectrum (<3.5 ppm) appear the peaks of the aliphatic side chains, including the very intense signals from the methyl groups (between 2 and −0.5 ppm). The protein was dissolved in 9 : 1 H_2O/D_2O, pH 7.0. The water signal was suppressed during acquisition and further reduced during processing.

18.1.3 Two-Dimensional NMR Spectroscopy

General Scheme of a 2D Experiment The interpretation of a 1D spectrum is almost impossible for more complex molecules due to spectral overlap (Figure 18.12). We will now illustrate on the basis of a 2D experiment (Figure 18.13) how additional dimensions can help to resolve spectral overlap. In addition to the building blocks (preparation and detection) known from the 1D experiment, two new blocks are introduced: the indirect evolution time t_1 and the mixing period.

After the preparation period, the spins precess freely throughout the fixed time t_1, in which the magnetization is "encoded" with the chemical shift of the first nucleus. During the following mixing period the state of the magnetization at the end of t_1 is read out, and magnetization is transferred from the first nucleus to another one. In general, scalar coupling and dipolar interactions (Sections 18.1.3, The COSY Spectrum, The TOCSY Spectrum, and The NOESY Spectrum) are used to transfer magnetization during the mixing period. It can consist of a single pulse (as in case of the COSY experiment), or of several pulses and delay times for more advanced experiments. The data acquisition (t_2 time, also called direct evolution time), during which the magnetization is encoded with the chemical shift of the second nucleus, completes the experiment.

Fourier transformation with respect to t_2 produces a conventional 1D spectrum that represents a snapshot at a given time t_1. If one records several individual experiments, where in each case the indirect time is increased by a fixed amount Δt_1 (Figure 18.14a), then in analogy to a movie this series of experiments displays the time development of the spin system during t_1 (Figure 18.14b and c). Hence, additional Fourier transformation with respect to t_1 produces the final 2D spectrum (Figure 18.14d, left-hand hand side), which is commonly shown in contour mode (Figure 18.14d, right-hand side). The projections of this 2D spectrum on both frequency axes result in the conventional 1D spectra.

The appearance of such a 2D spectrum depends on the frequencies recorded during t_1 and t_2. In a heteronuclear 2D spectrum two different types of nuclei are detected in both frequency dimensions, whereas in the homonuclear case the same type of nucleus is detected. Thus, the homonuclear 2D spectrum is symmetrical with a diagonal line crossing the spectrum (Figure 18.15). All additional peaks lying symmetrical to this diagonal are called cross-peaks.

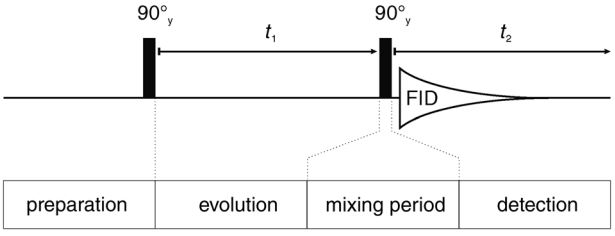


Figure 18.13 Schematic illustration of a two-dimensional NMR experiment. As a prototypical example the COSY experiment is shown. Besides the preparation and detection period known from the 1D experiment, the 2D experiment includes additional evolution and mixing periods. The mixing period serves to create correlations between interacting spins, which is facilitated by a single 90°-pulse in the COSY experiment. During the evolution and detection periods the system evolves during the t_1 and t_2 times.

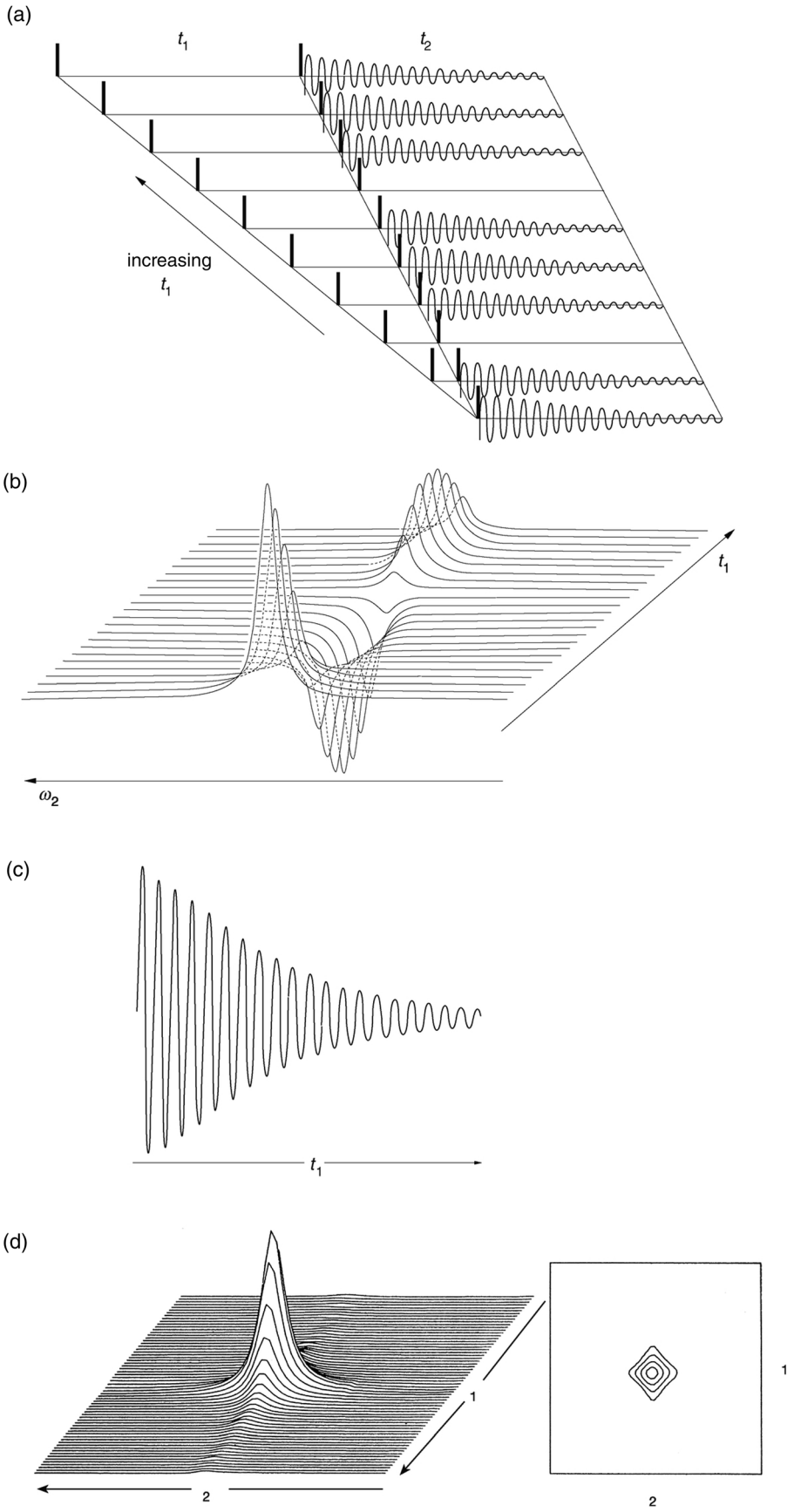


Figure 18.14 Origin of a 2D NMR spectrum through two-dimensional Fourier transformation of the acquisition data. (a) Between consecutive 1D experiments of a 2D spectrum the time t_1 is incremented, thus stepwise scanning the indirect time-domain, Source: adapted from: Cavanagh, J., Fairbrother, W.J., Palmer, A.G. III, and Skelton, N. (1996) *Protein NMR Spectroscopy*, Academic Press. (b) Fourier transformation with respect to t_2 results in a series of one-dimensional spectra modulated in t_1. (c) Intersection through the series of experiments from part (b). The intersection through the respective peak maxima parallel to t_1 corresponds to a FID in the indirect time-domain. (d) Fourier transformation with respect to t_1 produces a two-dimensional absorption line. The left-hand image shows a perspective view of this line; the right-hand image displays the conventional contour plot with height lines. Source: parts (b)–(d) adapted from Derome, A.E. (1987) *Modern NMR Techniques for Chemistry Research,* Pergamon.

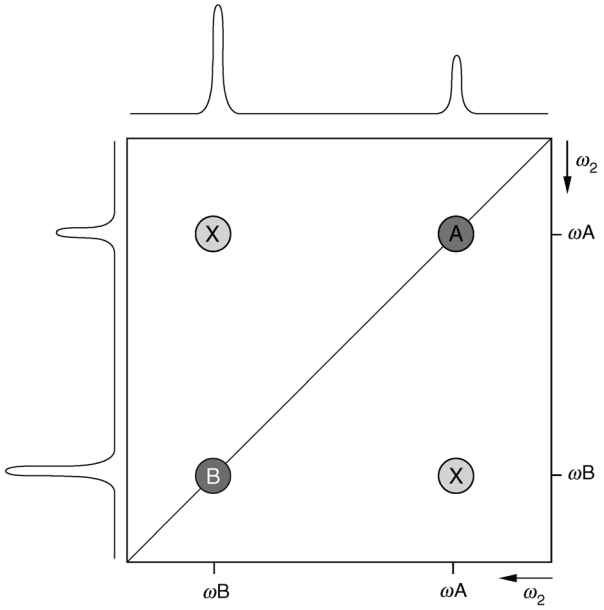


Figure 18.15 Schematic illustration of a homonuclear 2D NMR spectrum for two interacting spins A and B. A and B mark the diagonal peaks of both spins, which appear at the same frequency in both dimensions. X denotes the cross-peaks between the spins that appear at different frequencies in each dimension. The cross-peaks arise from magnetization transfer between both spins A and B during the mixing period of the experiment. The diagonal peaks arise from parts of the magnetization that remain on the same spin after the mixing period. The projections of the spectrum on both axes as well as the diagonal correspond to conventional 1D spectra.

The diagonal results from parts of the magnetizations that remain on the same spin after the mixing period (same frequency ω_1, ω_2 in both dimensions), that is, from parts that were on the same nucleus during both evolution times. Therefore, the diagonal corresponds to a conventional 1D spectrum. In contrast, cross-peaks correlate two signals from two nuclei that have exchanged magnetization during the mixing period (nucleus A with frequency ω_A and nucleus B with frequency ω_B). Thus, they indicate an interaction between the two nuclei, which constitutes the relevant information of the 2D spectrum.

Three homonuclear 2D techniques are of major importance for the NMR spectroscopy of small proteins (<10 kDa): COSY, TOCSY, and NOESY (including the related ROESY experiment). All three spectra in combination facilitate the assignment of the individual resonance frequencies to the respective protons in the protein. The NOESY experiment additionally provides the necessary information (distances between protons) to calculate the three-dimensional structure.

The COSY Spectrum The COSY (*co*rrelation *s*pectroscop*y*) experiment creates proton correlations through magnetization transfer via scalar coupling. However, only protons connected through two- (^2J) or three-bond (^3J) couplings give rise to cross-peaks (Figure 18.16). No cross-peaks are observed for protons connected through ^4J and higher couplings because the respective constants for aliphatic compounds are almost equal to zero. Of particular importance for protein structure determination are $^3J(\mathrm{H^N} - \mathrm{H^\alpha})$ coupling constants due to the information provided by the Karplus relationship about the backbone torsion angle ϕ (Figure 18.11).

The TOCSY Spectrum In the TOCSY (*to*tal *c*orrelation *s*pectroscop*y*) experiment a successive, multistep transfer via scalar J couplings distributes the magnetization over the whole spin system. Thus, TOCSY correlates all protons of a spin system (e.g., an amino acid) with each other. The obtained spectrum displays a characteristic peak pattern for each amino acid that facilitates the identification of the respective amino acid type. The peak patterns consist of several vertical rows of parallel running peaks that reflect the intraresidual correlations of the different proton pairs: for example, $\mathrm{H}^{\alpha}_{(i)} - \mathrm{H}^{\mathrm{N}}_{(i)}$, $\mathrm{H}^{\beta}_{(i)} - \mathrm{H}^{\mathrm{N}}_{(i)}$, $\mathrm{H}^{\gamma}_{(i)} - \mathrm{H}^{\mathrm{N}}_{(i)}$, and $\mathrm{H}^{\delta}_{(i)} - \mathrm{H}^{\mathrm{N}}_{(i)}$, respectively (Figure 18.16). However, the assignment of certain amino acid types (e.g., serine, cysteine, aspartate, asparagine, histidine, tryptophan, phenylalanine, and tyrosine) remains ambiguous because they display identical peak patterns due to a single methylene group as side chain spin system.

The NOESY Spectrum The NOESY (*n*uclear *O*verhauser *e*ffect *s*pectroscop*y*) experiment is one of the most important experiments for the structure determination of proteins because it provides distance information between individual nuclear spins. The correlations arise from the dipolar interaction between the spins, the *nuclear Overhauser effect*. The NOE acts through

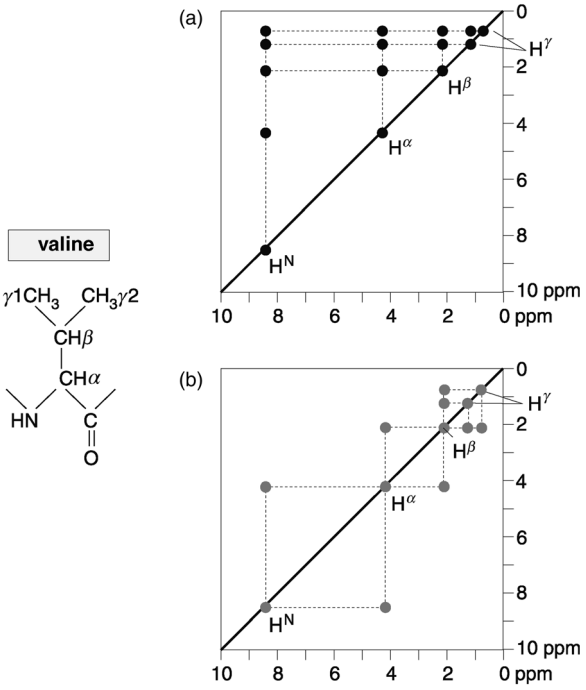


Figure 18.16 Representative peak patterns of the amino acid valine for a 2D TOCSY (a, black circles) and for a 2D COSY spectrum (b, grey circles). For clarity the intraresidue peaks below the diagonal of the TOCSY spectrum are omitted. An overlay of both spectra demonstrates that the peaks from the COSY spectrum also appear in the TOCSY spectrum. The peaks in the COSY spectrum result exclusively from protons correlated over three bonds. The structure of valine on the left shows the common nomenclature for protons in NMR spectroscopy.

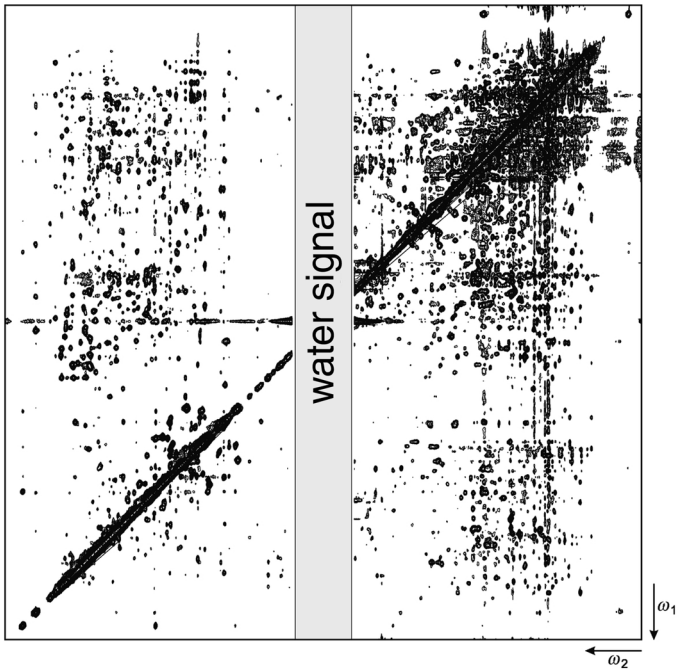
space and its strength is to a first approximation proportional to $1/r^6$ (where r is the distance between the interacting nuclei). Because the correlation of two nuclei depends on their spatial distance, magnetization transfer between two nuclei occurs only if their distance is less then ~5 Å. The NOESY experiment (Figure 18.17), therefore, provides correlations between all protons that are close enough in space. Even protons that are very distant in the protein sequence can give rise to cross-peaks if the tertiary structure brings them into close contact.

Homonuclear 2D NMR Experiments of Proteins: Limitations of Application The NMR spectroscopic analysis of larger proteins (>10 kDa) with 2D techniques is associated with several problems that make a detailed analysis difficult. The number of signals in two-dimensional ^{1}H NMR spectra increases disproportionally with the size of the protein. This causes significant peak overlap (especially in the aliphatic parts of the spectra), severely complicating the unambiguous assignment. In addition, the transverse relaxation time constants (T_2) become relatively short due to the slow tumbling of the molecules. The resulting line broadening of the peaks further enhances the spectral overlap. A second effect of shorter T_2 time constants is noticeable for all spectra that rely on scalar spin couplings to transfer transverse magnetization (COSY, TOCSY). During the transfer period, the transverse magnetization can become completely dephased due to T_2 relaxation – that is, the observable signal decays almost completely during the pulse sequence resulting in significant sensitivity losses.

Therefore, 2D TOCSY and 2D COSY are unsuitable for identifying the spin systems in larger proteins due to severe spectral crowding. The signal overlap further hinders the sequence-specific assignment (Section 18.1.5) based on NOE contacts, limiting the use of 2D NOESY experiments. These problems can be resolved by heteronuclear NMR experiments as well as multidimensional (>2D) NMR spectroscopy.

Heteronuclear NMR Experiments Besides protons, biomolecules contain other magnetically active nuclei (the so-called heteronuclei). Multi-dimensional NMR experiments for structure determination rely particularly on the magnetically active isotopes of carbon (^{13}C) and nitrogen (^{15}N). Due to the low natural abundances and small gyromagnetic ratios (Table 18.1), the relative sensitivities of ^{13}C and ^{15}N are low compared with protons. Therefore, to increase the sensitivity of heteronuclear experiments, two general strategies exist:

- First, recombinant expression of proteins in bacteria allows for the production of isotopically enriched proteins. To this end, bacteria are cultivated in minimal media, which contain



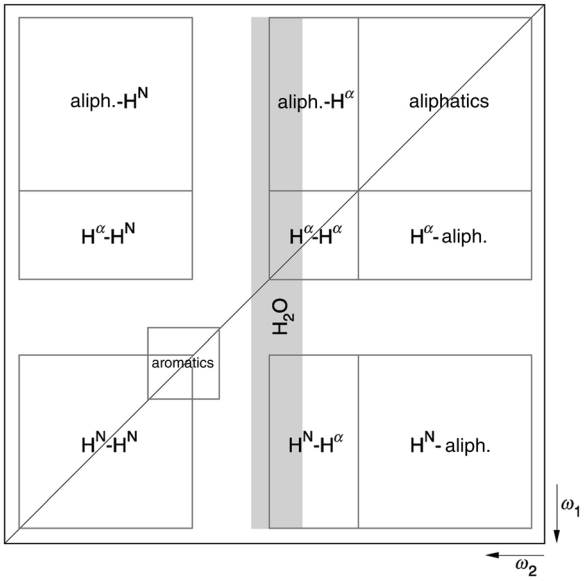


Figure 18.17 (a) Representative 2D NOESY spectrum of a 115 amino acid long protein. The very strong water signal in the center of the spectrum was removed during Fourier transformation. Grey rectangles in (b) schematically highlight the different spectral regions of this spectrum. For each rectangle the observable NOE signals in the respective region are given. The water line is shown as a grey bar.

$^{15}NH_4Cl$ as the sole nitrogen source. For enrichment of ^{13}C, the sole carbon source is ^{13}C-glucose. In this manner, singly-labeled (^{15}N or ^{13}C) or even doubly-labeled (^{13}C, ^{15}N)-samples are produced. Additionally, one can obtain deuterated proteins if D_2O instead of H_2O is used as the solvent for the culture medium.

- Second, the signal-to-noise ratio of an NMR experiment depends among other factors on the gyromagnetic ratios of the excited and detected nuclei. Therefore, the direct excitation/detection of heteronuclei is less sensitive relative to excitation/detection of protons. Thus, experiments in general rely on the transfer of the large magnetization of protons to an attached heteronucleus (and vice versa). This achieves an optimal signal-to-noise ratio at only minor magnetization losses. For historical reasons, such experiments are called inverse heteronuclear experiments.

HSQC – Heteronuclear Single Quantum Coherence The HSQC (heteronuclear single-quantum coherence) experiment constitutes the most important experiment that transfers

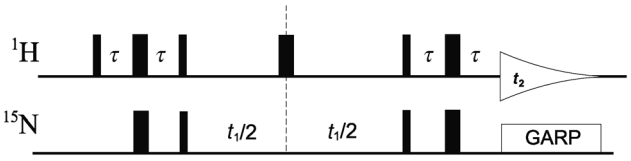


Figure 18.18 Pulse sequence of the HSQC experiment. Narrow black bars represent 90° pulses; broad black bars 180° pulses. The upper line displays pulses on the proton (^{1}H) frequency; the lower line on the nitrogen (^{15}N) frequency.

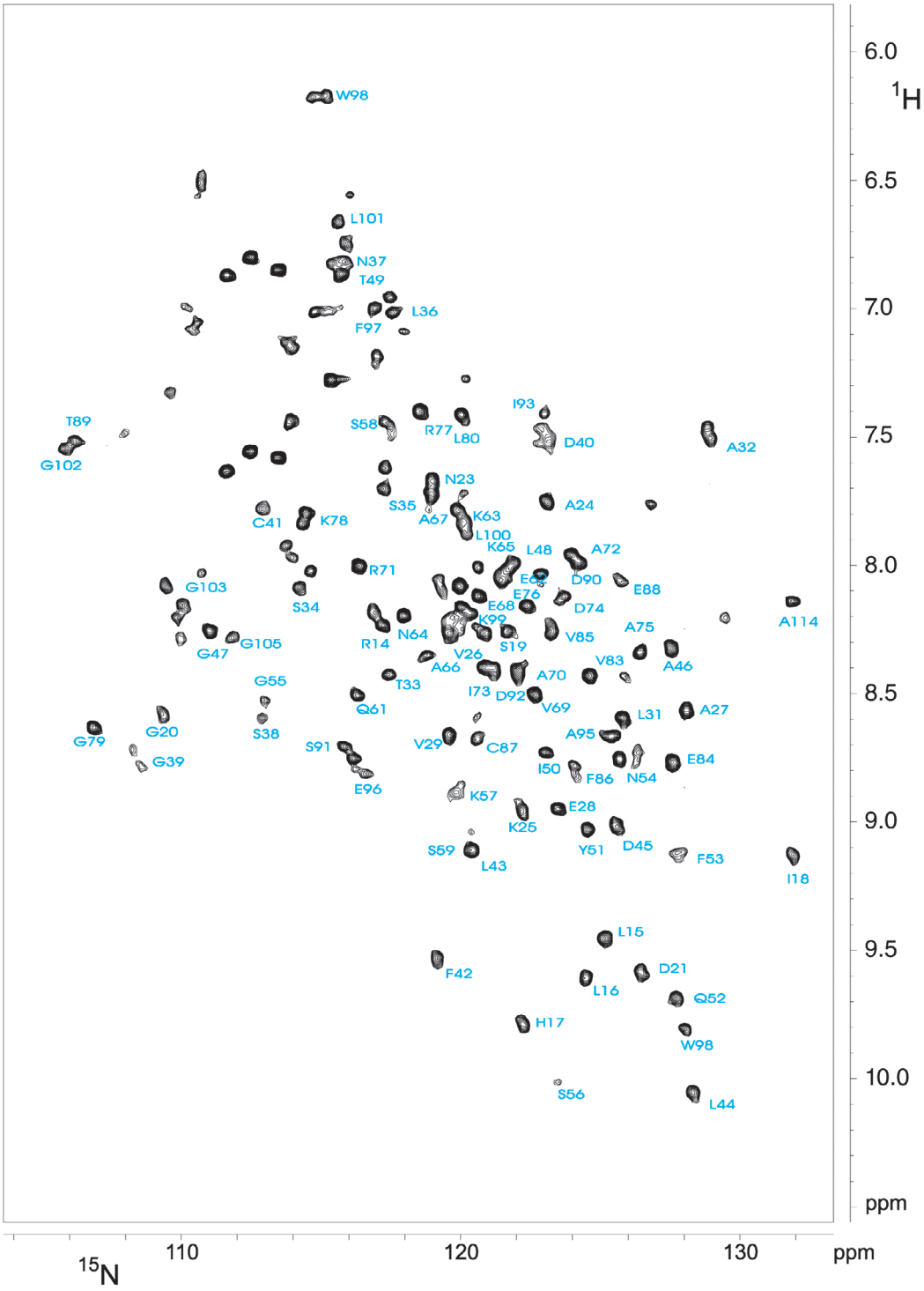


Figure 18.19 HSQC spectrum of severin DS111M at 32 °C and pH 7.0. The assignment for each peak is shown as single letter code (the numbers refer to the position in the protein sequence). The nitrogen frequency is plotted on the *x*-axis; the proton frequency on the *y*-axis.

magnetization to a heteronucleus and back (Figure 18.18). The HSQC correlates the frequency (ω_1) of a heteronucleus (^{13}C or ^{15}N) with that of the directly bound proton (ω_2). For example, in a two-dimensional ^{15}N-HSQC each peak represents one proton bound to one nitrogen atom, that is, the spectrum consists essentially of all the amide signals (H^N–N) of the protein backbone. Additionally, peaks arise for the aromatic, nitrogen-bound protons of the tryptophan and histidine side chains, and for the side chain amide groups of asparagine and glutamine, respectively (Figure 18.19). In the latter case, two peaks appear at the same nitrogen frequency because two amide protons are bound to the same side chain nitrogen. Under favorable conditions, the nitrogen-bound protons of the side chains of arginine and lysine are also visible. The advantage provided by the additional nitrogen dimension of the HSQC experiment is that it resolves amide proton resonances that often overlap in 1D and homonuclear 2D spectra of larger proteins. Compared with a homonuclear spectrum, the HSQC has of course no diagonal because it correlates completely different types of nuclei during the t_1 and t_2 times.

18.1.4 Three-Dimensional NMR Spectroscopy

The modularity of NMR experiments opens up the possibility for multidimensional NMR simply by introducing further dimensions. For example, we can create a 3D experiment through replacement of the acquisition time after the first mixing period of the 2D experiment (Figure 18.13) with another indirect evolution time and subsequent second mixing period (Figure 18.20). In four-dimensional NMR, a third indirect time follows as well as an additional mixing period. The different indirect times are each incremented individually. The direct data acquisition forms the end of each multidimensional experiment. We start our discussion of 3D NMR with the pulse sequences that are combinations of two 2D experiments because they are conceptually easier. Later, we will describe the so-called triple-resonance experiments that correlate three different types of nuclei (^{1}H, ^{13}C, ^{15}N).

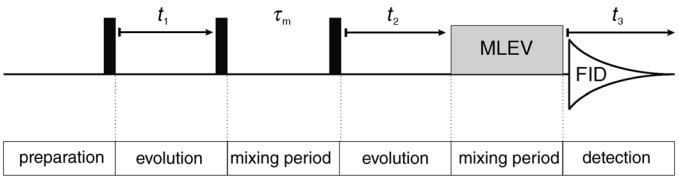


Figure 18.20 Schematic illustration of a 3D experiment showing the NOESY-TOCSY experiment as an example. Compared with the 2D experiment, a 3D experiment contains additional evolution times and mixing periods. The mixing period of the NOESY transfer step consists of two pulses with a delay time (τ_m) in between; the mixing period of the TOCSY transfer step consists of a complicated series of pulses called the MLEV mixing sequence (after its inventor Michael Levitt).

The NOESY-HSQC and TOCSY-HSQC Experiments We have mentioned above that spectral overlap limits the application of 2D spectra (NOESY or TOCSY) for larger proteins. Due to the dispersion of the peaks in a cube instead of a plane, the introduction of a third dimension can greatly resolve this overlap. In general, a heteronuclear coordinate like ^{15}N or ^{13}C constitutes the third (vertical) dimension of this cube because the wider frequency range of the heteronucleus provides a better signal resolution than an additional proton dimension.

We can create such a 3D experiment simply by combining the pulse sequences for a 2D NOESY and a 2D HSQC, in which instead of the data acquisition the HSQC experiment follows directly after the NOESY experiment. The created experiment is called ^{13}C- or ^{15}N-NOESY-HSQC. In an analogous way, we can convert a 2D TOCSY experiment into a 3D TOCSY-HSQC by combining a 2D TOCSY with a 2D HSQC. The ^{15}N-NOESY-HSQC and ^{15}N-TOCSY-HSQC represent the basic experiments for the sequence-specific assignment of medium-sized proteins (10–15 kDa). The respective ^{13}C variants are very useful in assigning the side chains and in identifying NOE signals between the side chain protons.

The HCCH-TOCSY and HCCH-COSY Experiments The HCCH-TOCSY and HCCH-COSY experiments are alternatives to the ^{15}N-TOCSY-HSQC, whose sensitivity strongly decreases for larger proteins. Both experiments transfer magnetization exclusively through scalar *J*-couplings between nuclei. For example, initially the magnetization transfer takes place from the H^α proton to the C^α nucleus (Figure 18.21). From there the magnetization transfer continues to the next carbon nucleus of the side chain, or in the case of the HCCH-TOCSY continues to all carbon nuclei of the side chain. Because the $^1J_{CC}$-coupling is about 35 Hz, the

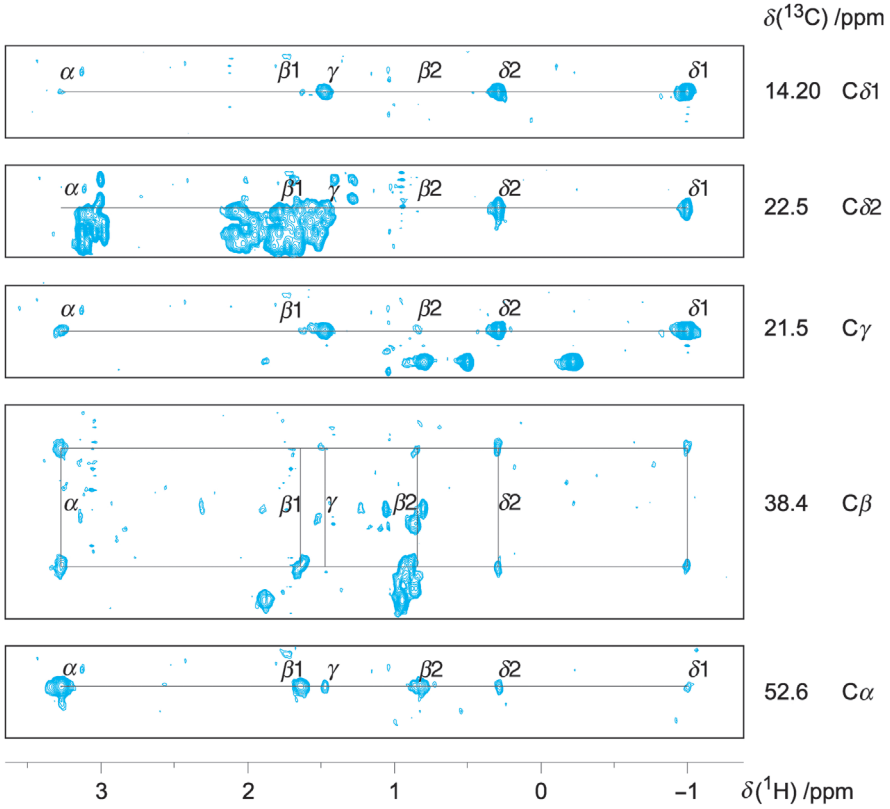


Figure 18.21 Slices of different planes from an HCCH-TOCSY experiment of reduced DsbA from *Escherichia coli*. All correlation signals of the amino acid Leu185 are shown. The ^{13}C chemical shift of the peaks is given next to the corresponding plane; the ^{1}H chemical shift is plotted horizontally along the lowest plane. Individual proton assignments are given next to the respective peak; carbon assignments are next to the slices of the respective plane. Collectively, these peaks produce the typical spin system pattern for a leucine residue also seen in a 2D TOCSY.

mixing time is shorter compared to the homonuclear case ($^{n}J_{HH} < 10$ Hz) and thus more sensitive for larger proteins. The time duration for the magnetization transfer is calculated according to $t_{mix} = 1/(2J)$. The data acquisition follows after the final magnetization transfer from each carbon nucleus to the directly bound proton. In general, the appearance of an HCCH-TOCSY (or HCCH-COSY) spectrum is identical to that of a ^{13}C-TOCSY-HSQC spectrum. Again characteristic spin system patterns similar to the 2D TOCSY and 2D COSY facilitate the identification of the amino acid type.

Triple-Resonance Experiments For proteins larger than 15 kDa, spectral crowding affects even the 3D NOESY-HSQC and TOCSY-HSQC spectra, thus challenging the protein backbone assignment (Section 18.1.5, Analysis of Heteronuclear 3D Spectra). The sequential assignment of large proteins therefore relies on triple-resonance experiments due to their simple appearance. For each amino acid only few peaks appear – often only one. Therefore, the problem with overlapping peaks occurs less frequently in triple-resonance spectra, and enables the sequential assignment of proteins up to 30 kDa. However, for certain nuclei the chemical shift of one residue may accidently match that of a different residue. This so-called degeneracy is especially common for the C^{α} nuclei. Identification of the correct connection between amino acids with degenerate chemical shifts is, therefore, one of the main obstacles for sequential assignment via triple-resonance experiments (Section 18.1.5, Sequential Assignment from Triple-Resonance Spectra). Because triple-resonance experiments correlate three different nuclei with each other, they require more expensive doubly-labeled ^{13}C, ^{15}N or triply-labeled ^{2}H, ^{13}C, ^{15}N protein samples.

A further advantage of triple-resonance experiments is their high sensitivity due to an efficient magnetization transfer through the strong ^{1}J- and ^{2}J-couplings (Figure 18.10) between the nuclei (i.e., directly via the covalent atom bonds). The required times for the transfer are, therefore, comparatively short so that relaxation losses are substantially decreased relative to a TOCSY experiment. This high sensitivity is another reason why signal assignment is possible for proteins up to a molecular weight of 30 kDa. For even larger proteins the sensitivity of triple-resonance experiments decreases mostly due to short transverse relaxation times. Especially the H^{α} protons enhance the relaxation of the C^{α} and H^{N} nuclei through dipolar interactions. Reducing the number of the aliphatic protons through protein deuteration greatly attenuates this

type of relaxation and extends the molecular weight of proteins that can be studied by NMR to 50 kDa and beyond (Section 18.1.7, Structure and Dynamics of High Molecular Weight Systems and Membrane Proteins).

Nomenclature of Triple-Resonance Experiments A large variety of triple-resonance experiments exists of which Figure 18.22 shows the most important representatives. Even though their nomenclature sounds cryptic, it is very descriptive. The name of the experiment specifies the detected nuclei, the magnetization transfer pathway, and the appearance of the spectrum. To this end, all nuclei through which magnetization is transferred are listed in a row. For example, the experiment denoted HNCO detects three nuclei (^{1}H, ^{13}C, ^{15}N) with the following flow of magnetization: $H^N_{(i)} \rightarrow N_{(i)} \rightarrow C'_{(i-1)}$ (Figure 18.22). Brackets in experiment names mark nuclei that serve only as "relay stations" and whose frequencies remain undetected. For example, the HN(CA)CO detects the same types of nuclei as the HNCO, however, the magnetization transfer differs: $H^N_{(i)} \rightarrow N_{(i)} \rightarrow C^{\alpha}_{(i)} \rightarrow C'_{(i)}$, (Figure 18.22). The C^{α} nucleus only relays magnetization from nitrogen to the carbonyl (C′) carbon, but its chemical shift is not recorded. Note that in both experiment types the magnetization follows the same pathway back to the amide proton for acquisition ("out-and-back" transfer). The appearance of the two spectra is similar. For each amino acid residue one peak arises that correlates the amide proton and nitrogen with a nitrogen-bound carbonyl carbon. However, while the HNCO spectrum shows the correlation with the C′ of the preceding residue (residue $i - 1$), the HN(CA)CO mainly shows the correlation with the intraresidual C′ (residue i).

The HNCA Experiment The HNCA constitutes one of the simplest and most useful examples of a triple-resonance experiment. The magnetization transfer pathway is given by $H^N_{(i)} \rightarrow N_{(i)}(t_1) \rightarrow C^{\alpha}_{(i)/(i-1)}(t_2)$; t_1 and t_2 specify the indirect dimensions during which the chemical shifts of the heteronuclei are encoded (Figure 18.20). The HNCA utilizes an "out-and-back" transfer to detect the amide proton chemical shift in t_3. In all cases the magnetization transfer occurs through strong J-couplings ($^1J_{HN} = 92$–95 Hz, $^1J_{NC} = 11$ Hz). Because the $^2J_{NC}$

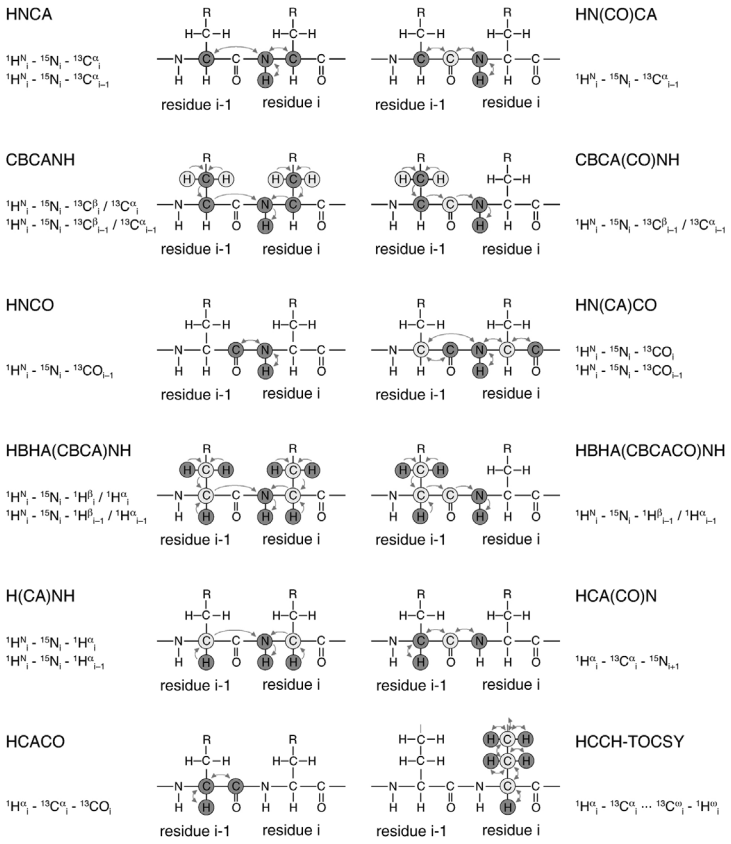


Figure 18.22 Overview of the most important triple-resonance experiments. Only the chemical shifts of the nuclei colored dark grey are detected during the experiment. Nuclei colored light grey serve as transmitters for the magnetization and remain undetected. Arrows mark the magnetization transfer pathway and direction. Under each experiment name all observable correlations are listed, in which subscripts i and $i - 1$ denote the position of the amino acid residues relative to each other. Even though the HCCH-TOCSY is not a triple-resonance experiment, it is included here due to its complementarity with the triple-resonance experiments for the assignment process.

coupling constant between the nitrogen nucleus and the C^{α} nucleus of the preceding amino acid is only marginally smaller (7 Hz) than the $^{1}J_{NC}$-coupling (11 Hz, Figure 18.10), magnetization transfer from nitrogen occurs to the C^{α} nuclei of both the amino acid it is part of and the preceding one. Therefore, for each amino acid two peaks arise in the HNCA spectrum: one intra- and one interresidue correlation. The related HN(CO)CA only shows the correlation to the preceding residue. In principle, the intra- and interresidue correlations of the HNCA enable the sequential assignment of the protein backbone. However, in practice one requires additional experiments to identify the cross-peaks of the preceding residue and also to resolve chemical shift degeneracies (Section 18.1.5, Sequential Assignment from Triple-Resonance Spectra).

18.1.5 Resonance Assignment

Sequential Assignment of Homonuclear Spectra NMR spectra contain all the necessary information about proton distances and torsion angles to calculate the structure of a protein or nucleic acid, which constitutes one possible aim of biomolecular NMR. To this end, it is necessary to assign each peak observed in the spectrum to the respective proton in the protein. Due to the large number of peaks in COSY, TOCSY, and NOESY spectra, this requires a simple and universal analysis method. Kurt Wüthrich (Nobel Prize in Chemistry 2002) and colleagues developed a method that could exactly achieve this – the sequence-specific assignment.

This method exploits the distance information contained in the NOESY spectrum. Because of their direct proximity, an amino acid $i+1$ displays specific contacts with amino acid i (in each case i denotes the relative position of the amino acids to each other). For example, due to the molecular geometry, the distances of the amide proton of residue $i+1$ to the H^{α}, H^{β}, or H^{γ} protons of amino acid i, respectively, are almost always less than 5 Å (Figure 18.23). Therefore, at the amide proton chemical shift of amino acid $i+1$ (horizontal frequency axis) cross-peaks arise at the chemical shifts (vertical frequency axis) of the respective protons of amino acid i. These interresidual peaks between neighboring amino acids are also called *sequential* peaks.

The method of sequence-specific peak assignment requires a distinction between the inter- and intraresidual peaks for a given amide proton chemical shift. A simple comparison of the 2D NOESY spectrum overlaid with the 2D TOCSY spectrum of the same sample facilitates this distinction (Figure 18.23). While the sequential cross-peaks provide information about the connectivity to the preceding amino acid, the characteristic pattern of the intraresidual cross-peaks determines the amino acid type.

Prolines interrupt this chain of sequential connectivities because they lack an amide proton and therefore show no peak in the amide region of the spectrum. However, proline residues, which adopt the more frequent *trans*-conformation (Figure 18.24), give rise to sequential $H^{\alpha}_{(i-1)} - H^{\delta}_{(i)}$ cross-peaks at the H^{α} chemical shift or $H^{N}_{(i-1)} - H^{\delta}_{(i)}$ cross-peaks at the amide proton chemical shift of the preceding residue. Another problem, which arises especially for larger proteins, is that the large number of peaks often results in peak overlap, preventing an unambiguous continuation of the sequential assignment at some positions in the sequence.

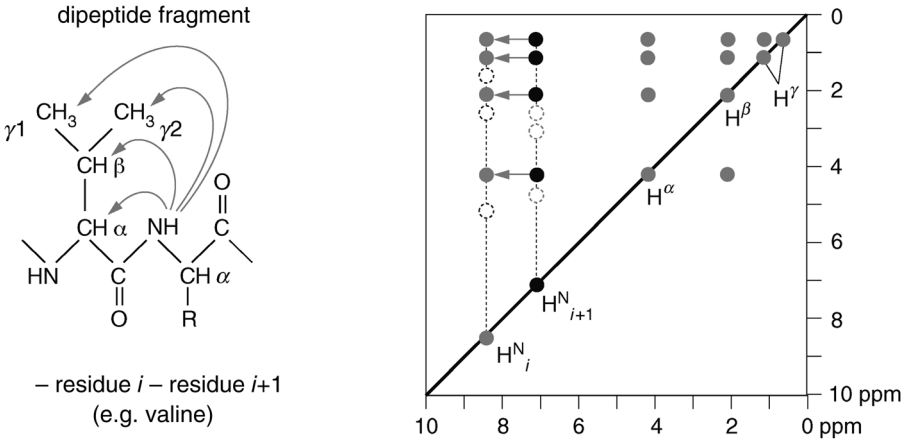


Figure 18.23 Overlay of a 2D NOESY (black circles) and a 2D TOCSY spectrum (grey circles) illustrating the schematic peak pattern of two neighboring amino acids. The interresidual, sequential NOESY peaks (filled black circles) on the succeeding amino acid $i+1$ form the basis for the sequence-specific assignment. Arrows in the dipeptide on the left-hand side mark the respective protons that give rise to the sequential cross-peaks. For completeness, dashed, open circles signify the intraresidual peaks of amino acid $i+1$ and the sequential peaks of amino acid $i-1$ at the amide proton chemical shift of amino acid i, respectively. For clarity, the symmetrical peaks occurring below the diagonal are omitted.

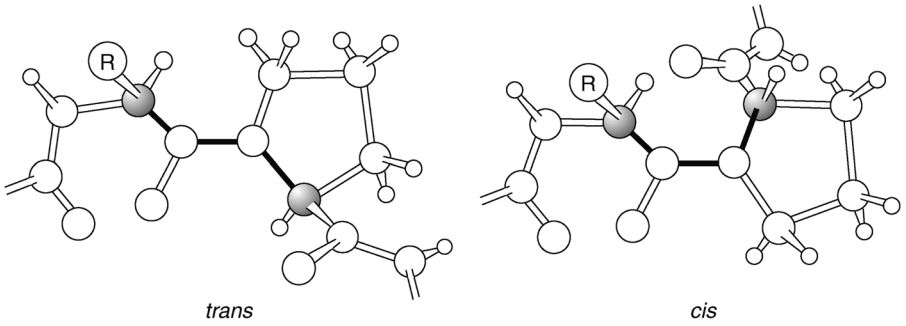


Figure 18.24 Conformation of the *cis*/*trans*-isomers of the peptide bond for an amino acid X and a proline residue. The C^{α} atoms of both residues as well as the connecting bonds that define the torsion angle ω are emphasized.

The first step of the sequence-specific assignment is to identify individual amino acids that serve as starting points for the assignment. Initially, the search is restricted to amino acids such as glycine, alanine, valine, or isoleucine because their characteristic peak pattern differs from other amino acid types. For example, glycine possesses two H^{α} protons. Detection of these two H^{α} peaks at the amide proton chemical shift and the appearance of the corresponding $H^{\alpha 1}$–$H^{\alpha 2}$ provide definite evidence for glycine (Figure 18.25). The characteristic double peak row of the methyl groups at 0–1.5 ppm (Figure 18.25) signifies valine, leucine, and isoleucine. In addition to the information provided by the amide region of the spectrum, the aliphatic region contains various diagnostic cross-peaks, especially for prolines, that help to validate the spin systems.

The next step is to identify the specific sequential contacts in the 2D NOESY spectrum and thus to determine the preceding residue for each of these distinct amino acids. Then, one determines for every newly identified residue the preceding amino acid in an iterative manner. Thus, the initially identified dipeptide is extended into an oligopeptide chain. The information about the amino acid types contained in these fragments enables one to place the fragments into the protein sequence – that is to assign each spin system to the respective amino acid.

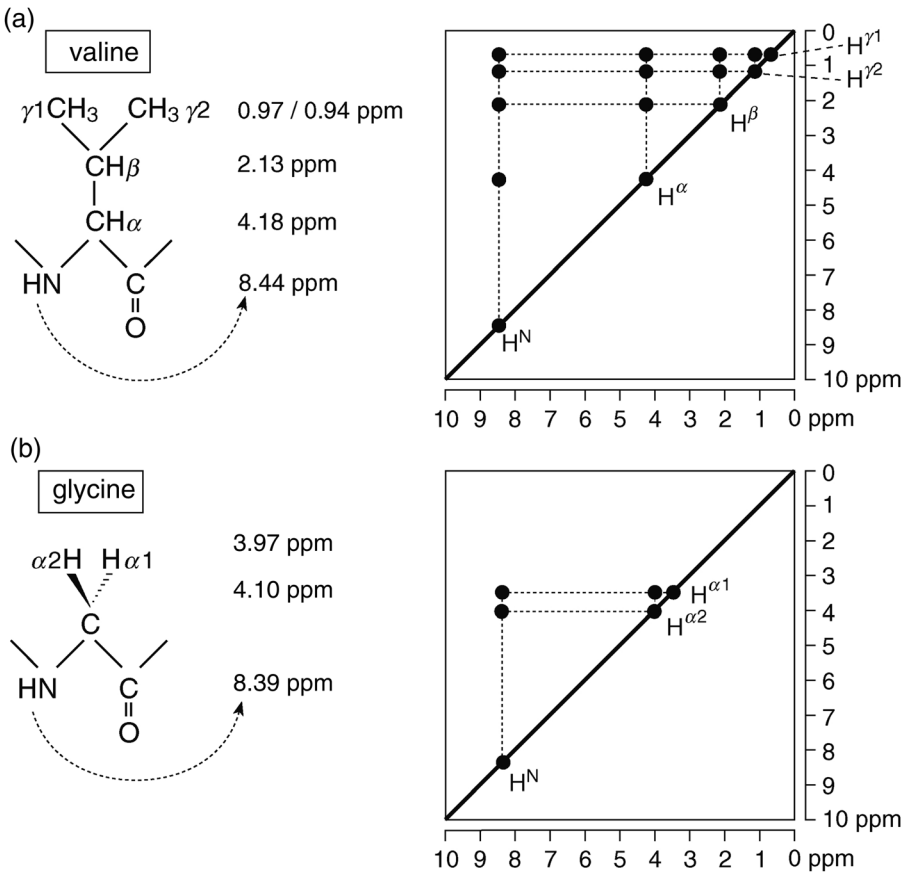


Figure 18.25 Schematic illustration of the characteristic peak patterns of (a) valine and (b) glycine. Both residues serve as starting points for the sequence-specific assignment due to their easily recognizable and distinctive patterns in the TOCSY spectrum. The left-hand side shows the structures of the respective amino acids together with the proton designations and typical chemical shift values. For clarity, the symmetrical peaks occurring below the diagonal are omitted.

Analysis of Heteronuclear 3D Spectra The method of sequence-specific assignment explained above for homonuclear 2D TOCSY and NOESY spectra is also appropriate for the respective heteronuclear 3D spectra ^{15}N-NOESY-HSQC and ^{15}N-TOCSY-HSQC. Every ^{15}N-plane of the spectra contains the NOESY and TOCSY peaks of an amide proton bound to its nitrogen. Again a superposition of the respective 3D TOCSY- and 3D NOESY-HSQC spectra facilitates the distinction between intra- and interresidual correlations. The ^{15}N-plane (of the 3D ^{15}N-NOESY-HSQC) is, therefore, a kind of sub-spectrum of the respective 2D spectrum (2D NOESY). A major difference from the 2D spectra is the frequency range sampled in the acquisition dimension of ^{15}N-edited 3D NOESY and TOCSY spectra. The experiments select only correlations to protons bound to ^{15}N-nuclei. Therefore, every 3D ^{15}N-NOESY-HSQC and ^{15}N-TOCSY-HSQC only contains frequencies between 12 and 5 ppm in the acquisition dimension; the side chain region beyond the water signal on the high field side of the spectrum is empty.

Selective Amino Acid Labeling ^{15}N-Labeling of selective amino acids constitutes an alternative way to determine the amino acid type. To this end, recombinant *Escherichia coli* cells are cultivated in a minimal medium containing all 20 naturally occurring amino acids. This medium composition greatly suppresses the cellular *de novo* synthesis of amino acids because the cells take up the amino acids directly from the medium. Addition of a commercially available 1-^{15}N-L-amino acid to the medium achieves the selective labeling of this residue type. All other amino acids are added to the medium as "normal" ^{14}N-amino acids. In particular, those unlabeled amino acids need to be added in excess, which can be metabolically synthesized from the selectively labeled amino acid (the so-called scrambling). An HSQC spectrum acquired on a selectively labeled protein shows only the signals from the labeled amino acid type. Therefore, the labeling of different amino acid types facilitates a quick assignment of the NMR signals and further validates the assignment obtained from TOCSY and NOESY spectra.

In general, selective amino acid labeling is less expensive than the production of doubly-labeled proteins because it requires less protein amounts and the ^{15}N-labeled amino acids are moderately priced (the actual price depends on the specific amino acid type). However, to assign all residues in a protein, different selectively labeled proteins have to be made, which increases the workload compared to the production of a single, uniformly labeled sample. Thus, the production of multiple samples can easily compensate for the previous savings.

Sequential Assignment from Triple-Resonance Spectra To illustrate the sequential assignment with triple-resonance experiments, we will restrict ourselves to the most popular

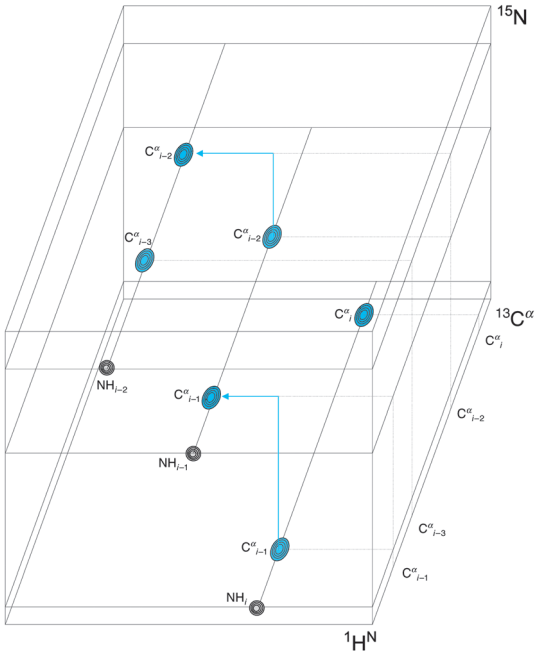


Figure 18.26 Schematic illustration of the sequence-specific assignment for an HNCA spectrum. The ^{1}H–^{15}N projection of the 3D spectrum shows one peak for each amino acid (grey peaks). At each amide proton chemical shift two cross-peaks exist, one originating from the correlation with the Cᵅ nucleus of the residue it is part of and one from the correlation with the preceding residue (blue). Starting from the blue signals, one can "walk" via the blue signals stepwise through the amino acid sequence (blue arrows).

representatives. The HNCA spectrum will serve again as an example to explain the general appearance of triple-resonance spectra and the assignment strategy.

We have already shown in Section 18.1.4 (Triple-Resonance Experiments) that the HNCA spectrum consists of three frequency axes (^{1}H, ^{15}N, and ^{13}C). Its $^{1}H^{N}$–^{15}N projection looks like an HSQC spectrum and every peak in this projection corresponds to a single residue. As mentioned earlier, two cross-peaks appear in the $^{13}C^{\alpha}$ dimension at the chemical shift of each amide proton: one intraresidual correlation and one interresidual correlation to the $^{13}C^{\alpha}$ nucleus of the preceding amino acid (Figure 18.26). In principle, this sequential information of the cross-peaks suffices to "walk" through the complete sequence (Figure 18.27).

The "sequential walk" requires a clear distinction between intraresidual and sequential cross-peaks. In general, the intraresidue peaks tend to be slightly more intense than the interresidue peaks (due to the more efficient magnetization transfer via the $^{1}J_{NC}$ coupling). This tendency is, however, fallible because processes like relaxation influence the peak intensity. In contrast, targeted experiments like the HN(CO)CA force the magnetization through the carbonyl carbon and reject the intraresidual pathway, while the spectrum otherwise looks identical to the HNCA. Thus, one obtains exclusively the sequential cross-peaks and resolves the ambiguity that is present in the HNCA.

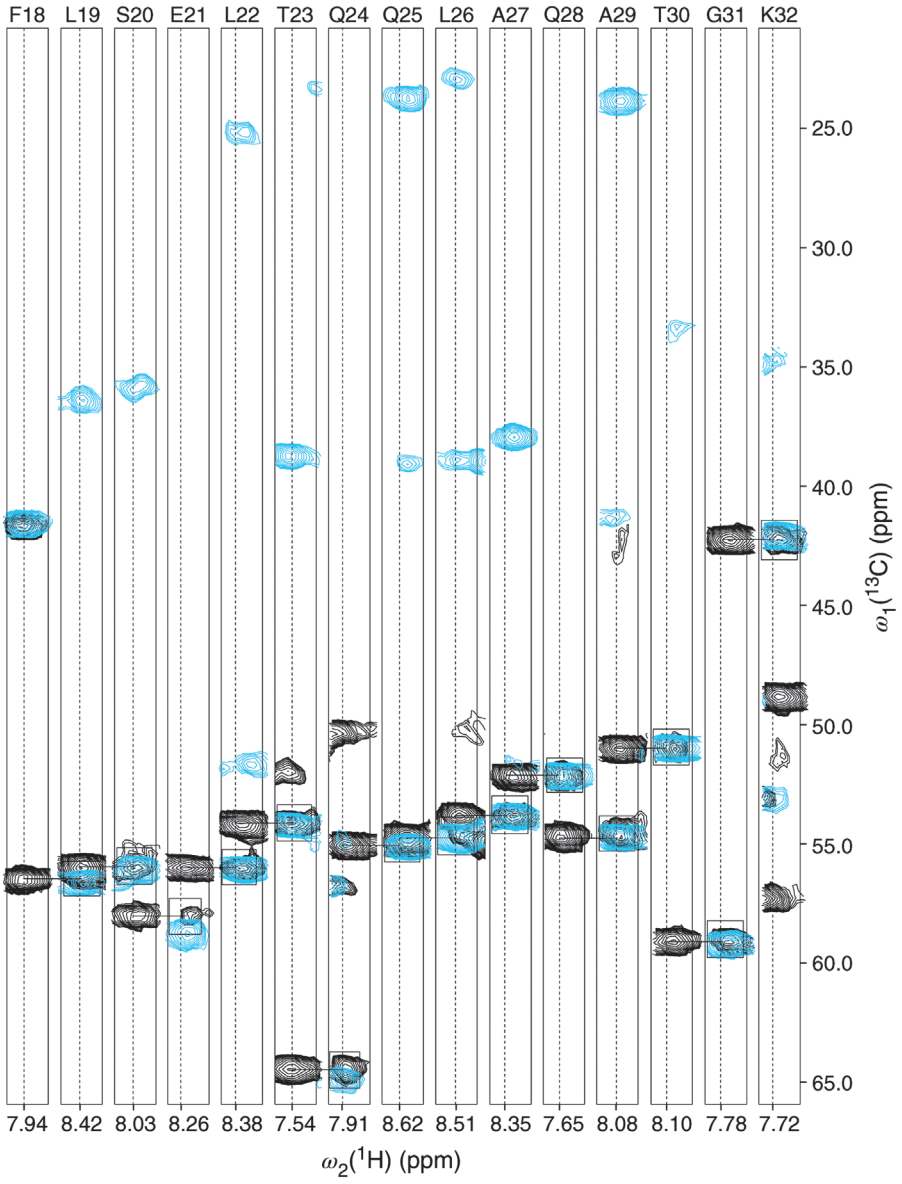


Figure 18.27 Slices from an HNCA (black contours) and a CBCA(CO)NH spectrum (blue contours) of huMIF. Each stripe corresponds to one amino acid from Phe18 to Lys32 at the respective amide proton (*x*-axis) and nitrogen chemical shifts (*z*-axis, not shown). The superposition of both spectra illustrates the "sequential walk" in which black horizontal lines indicate the sequential connectivities.

Proline residues and chemical shift degeneracies interrupt the assignment procedure (similar to the assignment of homonuclear spectra, Section 18.1.5 (Sequential Assignment of Homonuclear Spectra)). Proline residues lack an amide proton and therefore produce no cross-peaks, while chemical shift degeneracies prevent an unambiguous assignment of the preceding residues because several possibilities exist. As outlined below, other types of triple-resonance experiments correlating the amide group with other carbon nuclei can resolve the chemical shift degeneracy; however, proline residues will always interrupt the sequential assignment in amide proton-detected experiments.

Combination of the HN(CA)CO with the HNCO experiment constitutes an independent alternative to validate the connectivities found in the HNCA. The two experiments correlate the amide proton with the intraresidual carbonyl carbon (HN(CA)CO) or that of the preceding residue (HNCO), that is, the sequential assignment is established through C' instead of C^{α} connectivities. The superposition of both spectra results in a pattern analogous to the HNCA spectrum. Furthermore, the HNCO spectrum, which is the most sensitive triple-resonance experiment, can be used to resolve accidental signal degeneracies in the HSQC projection. In proteins, each amide proton–nitrogen pair is covalently attached to only one C'. Therefore, one observes one cross-peak per amide proton frequency. However, if one finds two cross-peaks at the frequency of an amide proton than this means that the signals of two amino acids are degenerate in the HSQC projection.

Two further pairs of experiments provide independent assignment strategies. The CBCANH and CBCA(CO)NH experiments give rise to intra- and interresidual correlations of the amide group with the C^{α} and C^{β} resonances, respectively. The closely related HBHA(CBCACO)NH and the HBHA(CBCA)NH experiments provide the correlations with the H^{α} and H^{β} resonances. The C^{α} and C^{β} chemical shifts obtained from the first two experiments are useful to narrow down the amino acid type. Because of the distinct value of the C^{α} and C^{β} chemical shifts, they enable a preliminary identification of alanine, glycine, isoleucine, proline, serine, threonine, and valine residues (Figure 18.28).

If the resonances of different amino acids are degenerate in the ^{15}N-HSQC projection, then the HCACO spectrum provides a simple alternative to differentiate these residues. This experiment establishes the correlations between the H^{α}, C^{α}, and C' frequencies of an amino acid and thus allows for the continuation of the assignment.

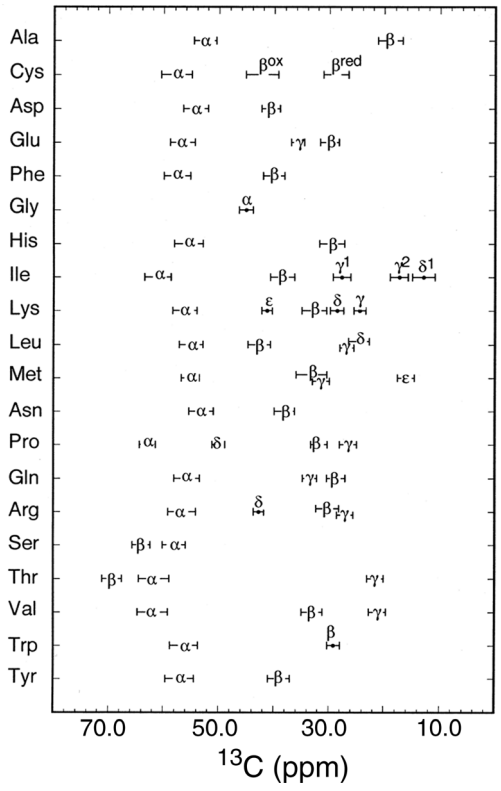


Figure 18.28 Typical ranges of C^{α} and C^{β} chemical shifts of the 20 amino acids. Source: adapted from: Cavanagh, J., Fairbrother, W.J., Palmer, A.G. III, and Skelton, N. (1996) *Protein NMR Spectroscopy*, Academic Press.

While triple-resonance spectra provide the connectivities between individual spin systems (and thus the sequential assignment) they only yield limited information about the side chain. The side chain assignment therefore relies on HCCH-TOCSY and the HCCH-COSY experiments, and starts from the known frequencies of C^{α} (obtained from the HNCA) or H^{α} (obtained from, for example, the ^{15}N-TOCSY-HSQC or the HCACO).

Summary: The general strategy to sequentially assign a protein by triple-resonance experiments requires the acquisition of several independent spectra. First, one establishes the sequential connectivity between the spin systems through at least two different nuclei (C^{α}, C^{β}, or C'). This strategy minimizes complications of the assignment by chemical shift degeneracies. Secondly, one restricts the amino acid type through the chemical shift information provided by CBCA(CO)NH or CBCANH spectra. Third, one assigns the side chains with HCCH-TOCSY and ^{13}C-NOESY-HSQC experiments (Section 18.1.4, The NOESY-HSQC and TOCSY-HSQC Experiments). Taking all the information together one can finally place the identified spin systems into the protein sequence.

18.1.6 Protein Structure Determination

Constraints for the Structure Calculation Up to now, we have described methods to identify and assign the observable NMR signals to the respective amino acids. Once the assignment of the resonances to individual nuclei has been completed, the next step is to extract structure-defining data from the NMR spectra. Most important are 2D NOESY, 3D ^{15}N-NOESY-HSQC, and 3D ^{13}C-NOESY-HSQC spectra (Section 18.1.4, The NOESY-HSQC and TOCSY-HSQC Experiments), as they provide a wealth of proton–proton distances. For a medium-sized protein (ca. 120 amino acids) one usually observes more than 1000 NOE contacts, which have to be assigned to specific protons on the basis of the previously established sequence-specific resonance assignment. Particularly important are the non-sequential NOEs that define the three-dimensional structure. *Medium-range* NOEs (less than four amino acids separate the residues involved in the NOE) provide information about the local backbone conformation of the protein and serve to determine secondary structural elements. *Long-range* NOEs (five or more amino acids separate the involved residues) define the relative orientation of the secondary structural elements to each other and are thus the essential parameters for the determination of the tertiary structure.

Because the NOE signal intensity I depends on the distance r between two nuclei i and j according to:

$$I(\mathrm{NOE}_{i,j}) \propto \frac{1}{r_{ij}^{6}} \tag{18.13}$$

the internuclear distances can be obtained through integration of the NOE signal. Alternatively, it is possible to qualitatively estimate their intensity. In both cases, however, the signal intensities have to be calibrated with respect to a NOE signal of known distance (e.g., to known distances in secondary structural elements). Depending on the signal intensities NOEs are classified into different distance groups with fixed distance boundaries (Table 18.2).

Table 18.2 Relationship between the intensity of a NOE signal and the respective proton distance. During structure calculations NOE distances act like elastic springs between the atoms. If an atomic distance exceeds the upper limit, then this NOE violation is penalized with an energy factor that depends on the extent of violation. This energy factor forces the respective protons closer together in the next iteration of the structure calculation. Because various effects can reduce the NOE intensity in the NMR spectrum irrespective of the atomic distance, usually no lower limits are used for the calculation; only the "normal" repulsion between the atoms due to their van der Waals radii acts.

NOE intensity	Distance (Å)	Upper limit (Å)
Strong	2.4	2.7
Medium	3.0	3.3
Weak	4.4	5.0
Very weak	5.2	6.0

In addition to distance restraints, structure calculations utilize the $^3J(H^N–H^\alpha)$ coupling constants provided by COSY or HNCA-J spectra (a variant of HNCA). As described in Section 18.1.2 (Spectral Parameters), $^3J(H^N–H^\alpha)$ coupling constants restrain the ϕ torsion angles of the protein backbone through the Karplus relationship (Figure 18.11).

Residual dipolar couplings (RDCs) represent a rather new class of NMR restraints that includes chemical shift anisotropy and cross-correlated relaxation. In isotropic solution rotational diffusion averages anisotropic interactions such as the dipolar coupling between two nuclear spins to zero. Dipolar couplings therefore do not result in line splitting in isotropic solution, but act as a relaxation mechanism. However, if the molecular tumbling is anisotropic and certain orientations are preferred or avoided (i.e., a molecular alignment is present), dipolar couplings average incompletely, resulting in scaled down values (relative to the static value) of the dipolar interaction. The corresponding couplings are called residual dipolar couplings. In contrast to NOEs and scalar couplings, which yield distance information about the local geometry, orientational restraints such as RDCs provide long distance restraints for the whole molecule. For example, RDCs determine the relative orientation of secondary structural elements or of individual protein domains to each other. Additionally, RDCs provide dynamic information about slow internal motions of a bond vector. The use of RDCs in structure calculations is, however, more complicated than for NOEs or *J*-couplings.

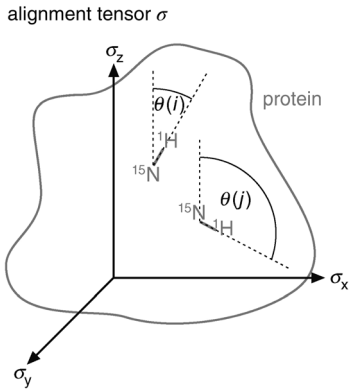


Figure 18.29 The orientation of an N–H bond vector *i* relative to the alignment tensor σ determines the magnitude and sign of the residual dipolar coupling. The alignment tensor is fixed to the protein; however, it depends in magnitude and orientation on the orienting medium (such as bicelles and Pf1).

All RDCs refer to the same reference frame that is fixed to the molecule – the anisotropic alignment tensor σ. Thus, RDCs depend on the orientation of the internuclear vector relative to the molecular reference frame, which is described by two angles (θ, ϕ) (Figure 18.29); θ defines the angle between the internuclear vector and the *z*-axis of the principle axis frame, and ϕ is the corresponding azimuthal angle. The principle axis frame is defined as that molecular frame in which the alignment tensor σ is diagonal (i.e., with principal values $|\sigma_z| > |\sigma_y| > |\sigma_x|$). The dipolar coupling $D^{A,B}$ between two nuclei A and B is given by:

$$D^{A,B} = 0.75 D^{A,B}_{max}\left[\left(3\cos^2\theta - 1\right)\sigma_z + \sin^2\theta\cos 2\phi\left(\sigma_x - \sigma_y\right)\right] \qquad (18.14)$$

The magnitude of RDCs scales with the degree of alignment. Experimentally, one prefers a weak alignment of the proteins to determine RDCs with a magnitude of a few Hertz. When the alignment is weak (corresponding to the alignment of 1 out of 1000 molecules) the spectra remain comparable to those in isotropic solution – that is the number of lines does not increase strongly and the NMR signals remain sharp. Several methods exist to achieve a weak alignment of a protein. Originally, particles were studied that aligned spontaneously in the magnetic field (through anisotropic tensors of the magnetic susceptibility). Meanwhile, researchers have resorted to systems that display a liquid-crystalline behavior in the magnetic field even when strongly diluted:

- bicelles (flat micelles consisting of two types of lipids with different acyl chain lengths);
- rod-like virus particles (tobacco mosaic virus or bacteriophage Pf1);
- poly(ethylene glycol)/alcohol mixtures;
- purple membranes of halophile bacteria.

In addition, mechanically stretched or compressed polyacrylamide gels also achieve a partial alignment of the solute (protein or nucleic acid). To determine RDCs, mostly heteronuclear experiments, which are also used for measurement of scalar couplings, are recorded.

Determination of the Secondary Structure Regular secondary structural elements like α-helices and β-sheets are regions with a defined conformation of the protein backbone, that is, with well-defined torsion angles and fixed proton–proton distances. Therefore, diagnostic NOE signal patterns and $^3J(H^N–H^\alpha)$ coupling constants allow for a distinction between the most important secondary structural elements (Figure 18.30). Characteristic for an α-helix are strong NOE signals between the amide protons $H^N_{(i)} - H^N_{(i+1)}$ and $H^N_{(i)} - H^N_{(i+2)}$, and also the very specific NOEs between $H^\alpha_{(i)} - H^N_{(i+3)}$ (Figure 18.30). $^3J(H^N\text{-}H^\alpha)$ coupling constants less than 5 Hz additionally confirm the existence of an α-helix. In contrast, for β-sheets a specific pattern of $H^\alpha_{(i)} - H^\alpha_{(j)}$ and $H^N_{(i)} - H^N_{(j)}$ NOEs arises between the two parallel or antiparallel strands of the sheet (*i* and *j* denote amino acids in different strands of the sheet). The extended structure of the protein backbone is further evident from strong $H^\alpha_{(i)} - H^N_{(i+1)}$ NOEs and $^3J(H^N–H^\alpha)$ coupling constants of more than 8 Hz for several consecutive residues.

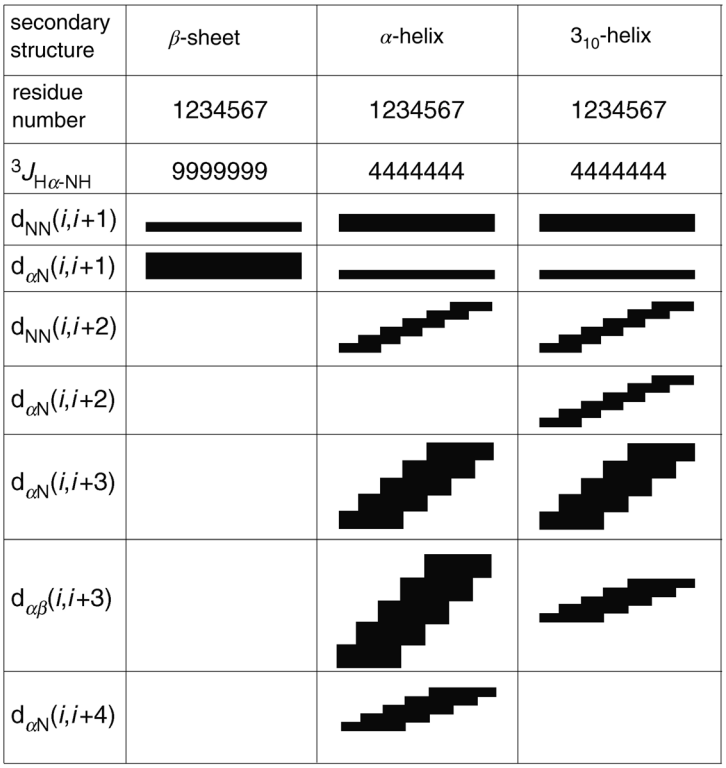


Figure 18.30 Characteristic NOEs and typical $^3J(H^N\text{–}H^\alpha)$ coupling constants for three regular secondary structures. The plot schematically illustrates typical intensities for the $H^N_{(i)} - H^N_{(i+1)}$, $H^\alpha_{(i)} - H^N_{(i+1)}$, $H^N_{(i)} - H^N_{(i+2)}$, $H^\alpha_{(i)} - H^N_{(i+2)}$, $H^\alpha_{(i)} - H^N_{(i+3)}$, $H^\alpha_{(i)} - H^\beta_{(i+3)}$, and $H^\alpha_{(i)} - H^N_{(i+4)}$ cross-peaks. The height of the rectangles reflects the peak intensity.

We will use the DS111M domain of severin to illustrate the identification of secondary structures. This second domain of severin plays a central role in the polymerization and depolymerization of filamentous actin, a constituent of the cytoskeleton. Severin DS111M consists of 114 amino acids and contains three α-helices evident from $H^\alpha_{(i)} - H^N_{(i+3)}$ connectivities (Figure 18.31). In addition, strong sequential $H^\alpha_{(i)} - H^N_{(i+1)}$ contacts and missing $H^N_{(i)} - H^N_{(i+1)}$ NOEs indicate the existence of five β-sheets. The $^3J(H^N\text{–}H^\alpha)$ coupling constants (>8 Hz) provide further support for a β-structure in these regions. Moreover, particularly strong $H^\alpha_{(i)} - H^\alpha_{(j)}$ contacts between two strands facilitate the identification of associated strands within the β-sheet (Figure 18.32). These short distances (~2.2 Å) are easily recognized in 2D NOESY or 3D ^{13}C-NOESY-HSQC spectra. Because the H^α resonances have chemical shifts close to that of water, it is advisable to record those NOESY spectra on samples dissolved in D_2O, in order to reduce the interference from the water signal.

Next to the direct structural information provided by distances, orientational restraints and torsion angles, hydrogen/deuterium (H/D) exchange rates of amide protons yield indirect structural information. Hydrogen bonds, which generally stabilize secondary structures, strongly attenuate the H/D exchange. Thus, slowly exchanging amide protons are often an indication for the existence of secondary structure (Figure 18.31). In addition, one obtains clues about the possible position of the residue in the protein structure. Amide protons in the center of the protein possess reduced exchange rates with the solvent due to their reduced accessibility, while amide protons at the surface display higher exchange rates.

To experimentally determine the exchange rate, the protein sample is freeze-dried and subsequently dissolved in pure D_2O. Over time, those signals disappear that are associated with rapidly exchanging amide protons. Conversely, signals of slowly exchanging amide protons remain visible in the spectrum for longer times (sometimes up to several months). Those amide protons localize almost exclusively to regions of the protein with regular secondary structure (Figure 18.31).

The chemical shifts of protein residues provide further evidence for secondary structure. Relative to the disordered state (the random coil shift) the chemical shift changes if an amino acid is in a secondary structural element. This difference is called the secondary chemical shift and reflects the uniformity of the chemical environment in regular secondary structure. For example, C^α or C' carbons experience downfield shifts (i.e., larger ppm values) in α-helices and

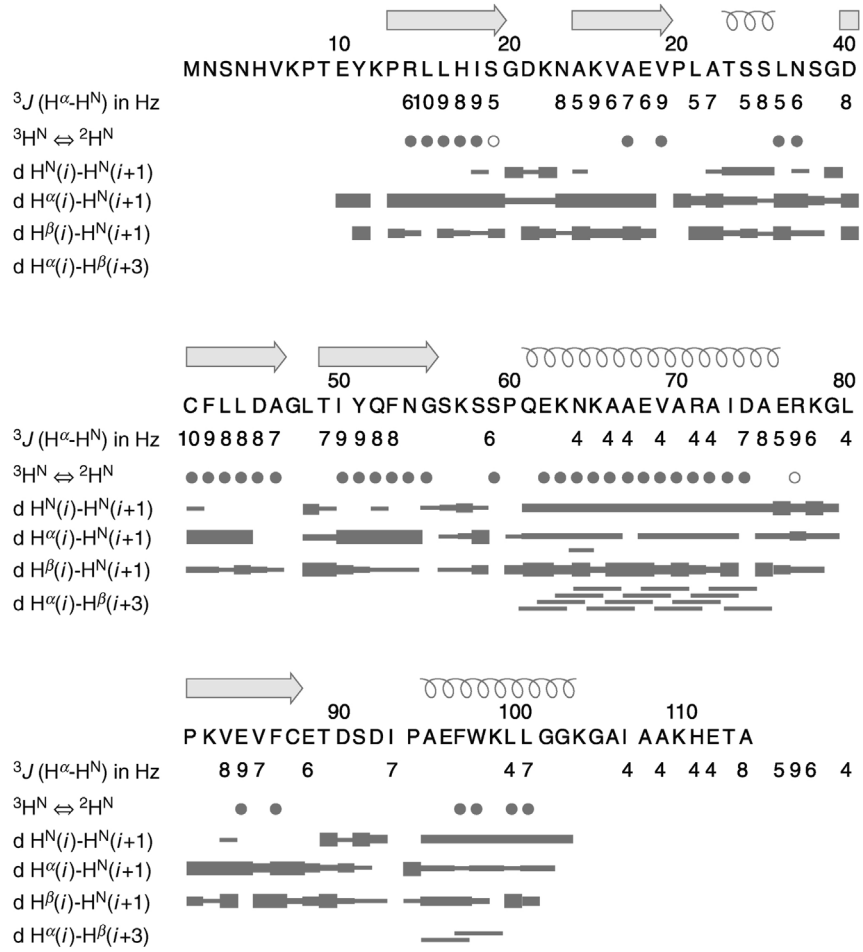


Figure 18.31 Overview of the sequential and short-range NOEs of DS111M. The height of the bars reflects the NOE peak intensity. The $^3J(H^N–H^\alpha)$ coupling constants are given below the amino acid sequence. Filled and open circles denote residues with slow and intermediate, respectively, amide proton exchange rates. Above the amino acid sequence the secondary structure is shown schematically (arrow = β-sheet, spring = α-helix).

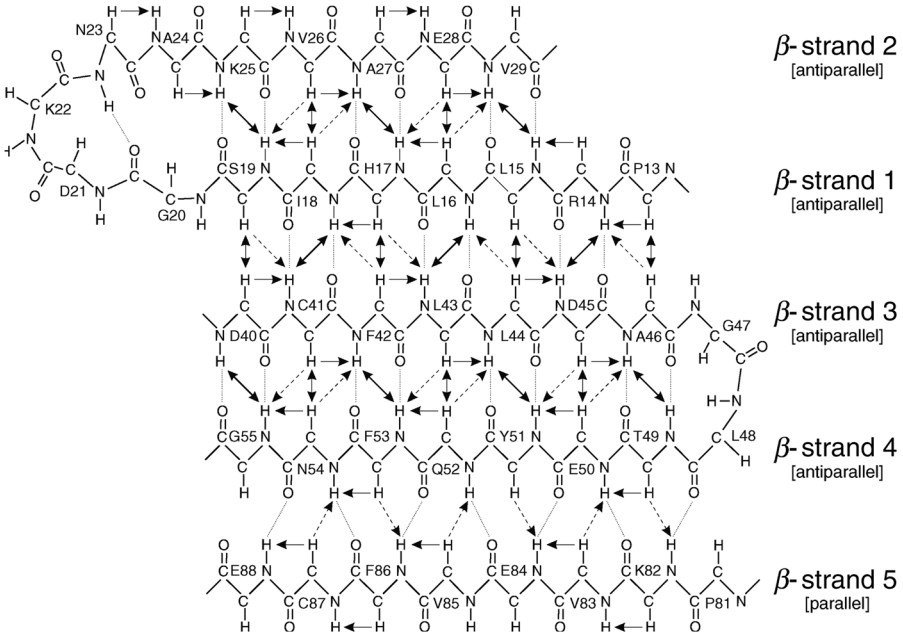


Figure 18.32 The complete network of NOE cross-peaks within the five-stranded β-sheet of severin DS111M. The four β-strands β1, β2, β3, and β4 run antiparallel to each other, while β-strands β4 and β5 run parallel. Arrows mark the NOE cross-peaks that occur within and between the backbones of the five β-strands. Solid arrows denote sequential $H^\alpha_{(i)}$-$H^N_{(i+1)}$ cross-peaks, while thick and thin double-headed arrows and dashed arrows denote interstrand $H^N_{(i)} – H^N_{(j)}$, $H^\alpha_{(i)} – H^\alpha_{(j)}$, and $H^\alpha_{(i)} – H^N_{(j)}$ cross-peaks, respectively. Additionally, thin dotted lines indicate hydrogen bonds.

upfield shifts (i.e., smaller ppm values) in β-sheets (vice versa for H^{α} and C^{β}). The method by which secondary structures are identified from secondary chemical shifts is known as the chemical shift index.

Therefore, already in the early stages of structure determination the above-described methods facilitate the identification of regions with secondary structure in the protein. However, the relative spatial positions of these elements to each other as well as the global fold of the protein remain unknown.

Calculation of the Tertiary Structure A computer-assisted structure calculation is used to convert the geometric data (distances and angles obtained from the analysis of the NMR spectra) into a three-dimensional structure of the biomolecule. However, the NMR data provide only a limited number of distance and angle restraints between atom pairs (preferentially the protons) that alone are insufficient to determine all atom positions. Luckily, this is only of minor importance because the bond geometries of many chemical groups are well known from X-ray diffraction experiments. This molecular information is contained in the so-called force field, which also includes further general atomic parameters such as van der Waals radii and electrostatic (partial) charges.

While NMR data are directly obtained from a target molecule and within error limits are considered as "real," force field parameters are derived from measurements on reference molecules. It is assumed that the force field parameters depend only on the chemical but not the spatial structure. Therefore, the force field constitutes a reasonable model for the real molecule. Because different possibilities exist to extract force field parameters from experimental reference data, several different force fields were developed for specific applications. Even though structure calculations depend on a force field, the experimental NMR data should determine the result irrespective of the choice of the force field. Additionally, compared to pure molecular dynamics simulations, NMR structure calculations use simplified force fields. For example, the solvent is generally considered only in the form of a fixed dielectric constant.

The most important degrees of freedom that determine the 3D structure of a protein constitute rotations about the $N–C^{\alpha}$ bond (torsion angle, ϕ) and the $C^{\alpha}–C'$ bond (torsion angle, ψ). To obtain reliable structures, NMR data should accurately define these angles and not the force field. Especially for loops at the surface, the N- and C-termini, or even the amino acid side chains this is often impossible due to their enhanced mobility. Thus, those regions do not assume a single defined structure, but rather fluctuate between different conformations. Flexible regions of biomolecules are therefore best described by an ensemble of conformations and a single static protein structure may not inevitably be the best representation of the microscopic reality. Practical experience shows that the number of distance restraints (NOEs) is more important than the accuracy with which the distances are determined. Thus, the distance classification introduced in Table 18.2 is sufficiently accurate for structure calculation.

In principle, two different methods exist to calculate the structure of a protein in solution (which also can be combined):

- The *distance geometry* method creates matrices from the NMR data and the force field that contain distance bounds for all atom pairs. Through mathematical optimization methods Cartesian coordinates are calculated for all atoms, which fulfill the distance bounds reasonably well. However, the solution to this problem is ambiguous and one can calculate many independent structures that all reasonably well agree with the NMR data. Because distance geometry takes the covalent geometry only insufficiently into account, all distance geometry structures require further refinement.
- *Simulated annealing* is a molecular dynamics (MD) method. Newton's equation of motion states that a force acting on an atom either accelerates or retards it. By numerically solving this equation, MD simulates the physical motion of atoms. NMR data are included as constraints and guarantee that the protein only adopts conformations that agree with the experimentally determined data. Starting from a template structure, a simulation period at high temperature allows the protein to find a structure that is compatible with the NMR data and the force field. Because at this stage both force field and experimental constraints are comparatively weak compared to the thermal energy, the simulated molecule can perform great conformational changes. The subsequent simulated annealing protocol decreases the simulation temperature and increases the impact of the force field and the experimental NMR constraints. In this way fluctuations in the structure are reduced until, finally, a 3D structure

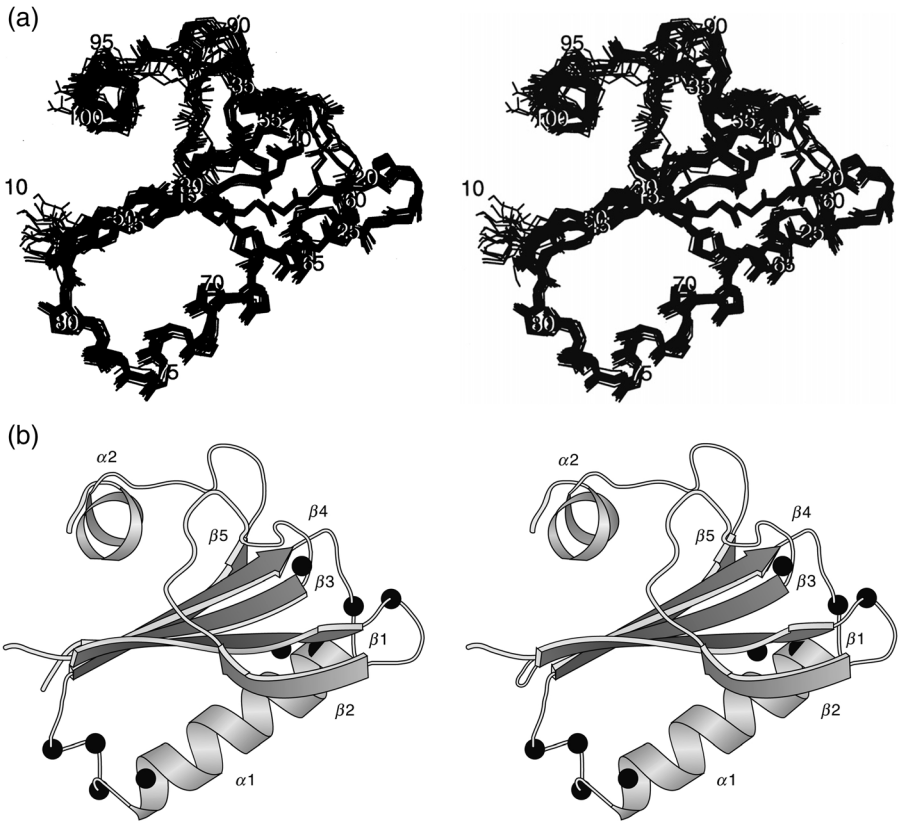


Figure 18.33 (a) Stereo image of the 20 lowest-energy structures of severin DS111M. Only the heavy atoms of the protein backbone (N, C^{α}, C′, and O) are shown. (b) Stereo image of the ribbon model of DS111M.

with minimal energy is obtained. Because the result of simulated annealing can depend on the starting structure, it is necessary to perform multiple calculations with different template structures.

NMR structure calculations do not provide a single structure but a family of conformers, which occupy a relatively narrow conformational space. The root-mean-square deviation (RMSD) expresses the variability within this structure family, in which small deviations indicate a narrow conformational space. In general, one determines the RMSD for each structure of this family relative to an average structure (which has to be calculated in advance). Alternatively, one can determine the RMSD by pairwise comparison of two structures of the family and calculating the mean of these deviations. The RMSD differs for individual parts of the protein structure because regions lacking a defined secondary structure show larger deviations due to few NMR constraints. Similarly, flexible regions with an increased internal mobility also give rise to higher RMSD values. Relaxation measurements (Section 18.1.7, Determination of Protein Dynamics) can clarify if the variability within the structural family is due to insufficient NMR restraints or is caused by local dynamics.

Figure 18.33 shows the result of a structure calculation for the protein severin DS111M. The individual conformers of the NMR ensemble are almost superimposable apart from a few poorly defined regions (the N-terminus and α-helix 2 at the C-terminus). Figure 18.33b displays DS111M as a ribbon model to enhance the presentation of the β-strands and α-helices. The orientation of the ribbon model is identical to the ensemble in panel (a). These structures were calculated with 1011 distance bounds and 55 ϕ torsion angles, and clearly demonstrate the necessity for a large number of constraints to obtain a well-defined structure.

18.1.7 Protein Structures and more — an Overview

NMR spectroscopy constitutes a versatile method with the possibility to investigate multiple chemical and biophysical problems apart from the mere determination of protein structures. The

determination of a protein structure is, thus, not the end but rather the start of a structural and biophysical characterization of a protein. Knowledge of the three-dimensional structure allows for a meaningful approach towards functional aspects, such as protein dynamics, interactions with other molecules (e.g., proteins, DNA, or ligands), catalysis mechanism, hydration, and protein folding. An extensive description of these techniques is beyond the scope of this chapter. The following subsections therefore only give a short overview of possible experiments and applications. The bibliography at the end of the chapter contains a selection of important textbooks and review articles, which explain individual subjects more precisely and extensively.

Speeding-up NMR Spectroscopy Due to the low sensitivity, NMR spectroscopy is a rather slow and time-consuming method. Even though simple 1D experiments last only a few seconds to minutes, the measurement time for more complex spectra increases dramatically with increasing dimensionality of the experiment. Two-dimensional experiments typically last for tens of minutes to hours; 3D experiments can take up to a few days to finish. Therefore, the acquisition of higher-dimensional spectra (4D or 5D), in which each indirect time dimension is incremented independently (as described above), is impractical.

Mainly two factors determine the length of an experiment, the relaxation delay (Section 18.1.2, The 1D Experiment) and the number of increments in the indirect dimension(s) (Section 18.1.3, General Scheme of a 2D Experiment). During the relaxation delay (ca 1–5 s) the magnetization recovers to its equilibrium value through T_1 relaxation. Hence, this delay determines how fast individual experiments (e.g., with different t_1 increments) can be repeated. The SOFAST- and BEST-type NMR experiments are specifically designed to selectively excite only a subset of the spins (usually the amide protons). The unperturbed protons enhance the longitudinal relaxation of the excited spins through dipole–dipole interactions. As a result one can reduce the relaxation delay to a few hundred microseconds and thus increase the repetition rate five to ten times.

The number of increments in the indirect dimension determines the achievable resolution in this dimension. Therefore, high-resolution spectra require more measurement time. Two different approaches exist to reduce the number of increments while maintaining the same resolution (or to increase the resolution in the same experiment time). In normal experiments the indirect time is incremented in constantly spaced intervals Δt_n. In contrast, fewer randomly-spaced increments in the indirect dimension (ca 30% relative to conventional sampling schemes) are used in the so-called non-uniform sampling method. Mathematical approaches (maximum entropy or compressed sensing), which are distinct from Fourier transformation, convert the sparsely sampled indirect time-domain data into a conventional spectrum.

The second approach relies on the simultaneous incrementation of at least two indirect time-domains to generate a two-dimensional projection of the *n*-dimensional spectrum. Imagine the three-dimensional cube of an HNCA spectrum with ^{1}H, ^{15}N, and ^{13}C on the *x*-, *y*-, and *z*-axes, respectively. Because of the co-incrementation of the ^{13}C and ^{15}N dimension, the resulting projection will intersect the ^{13}C–^{15}N plane at a certain angle with respect to the *z*-axis. Thus, in the projection the signals on the *y*-axis are the combination of the ^{13}C and ^{15}N frequencies, while frequencies on the unaffected *x*-axis correspond to the amide protons. Even though we are unable to imagine a four-dimensional object, mathematically it is very simple to transfer the described method to 4D experiments (and higher). Suitable projection reconstruction methods facilitate the creation of the *n*-dimensional spectrum from a limited number of projections. Alternatively, in automated projection spectroscopy (APSY) the chemical shifts for each peak are directly calculated from the projections without any reconstruction of the spectrum. For example, with projection spectroscopy one can acquire 6D experiments in three to four days.

Determination of Protein Dynamics Proteins are not rigid, static entities, but rather exist as ensembles of conformations. The internal motions, which give rise to transitions between different structural states, are collectively referred to as protein dynamics and occur on a wide range of time scales (ns–s). Protein dynamics play crucial roles in protein functions particularly for the interaction with binding partners, in enzyme catalysis, and in allosteric regulation. Luckily, many NMR spectroscopic parameters depend both on the mobility of the whole molecule (translational or rotational motions) and on internal motions (transitions between different conformations or bond rotations), providing insight into a wide variety of dynamic

processes ranging from fast fluctuations (lasting picoseconds) to slower conformational changes (lasting microseconds or more).

The measurement of relaxation parameters is possible for different nuclei. While ^{15}N relaxation data of the amide group provide information about the backbone flexibility of each amino acid, relaxation measurements of side chain groups (especially the methyl groups of valine, isoleucine, and leucine) determine their respective mobility. The ^{15}N-relaxation measurements have the advantage that they can be performed on inexpensive, ^{15}N-labeled protein. In addition, a simplified model can be used to analyze ^{15}N relaxation data because it is primarily the directly bound proton that determines the relaxation of the ^{15}N spin. In this model three parameters characterize the dynamic behavior: the rotational correlation time and the amplitude and the time scale of local motions. The correlation time τ_c describes the statistic motion of the whole molecule in the form of a rotational diffusion, which for proteins is on the order of several nanoseconds. The ^{15}N-relaxation measurements have been particularly successful in the identification and characterization of excited protein states that have populations as low as a few percent but which are important for protein function and misfolding.

Thermodynamics and Kinetics of Protein–Ligand Complexes NMR spectroscopy can elucidate many aspects of the interaction of proteins with small ligands, polypeptides and other proteins, or nucleic acids (DNA and RNA). It characterizes the dynamic, kinetic, and thermodynamic properties of protein–ligand complexes. Even without precise structural information about the ligand, it is possible to identify the residues of a protein involved in binding a ligand. To this end, ^{15}N- or ^{13}C-labeled protein is titrated with an NMR-invisible, unlabeled ligand and a HSQC is recorded at each titration point. Due to ligand binding, some peaks display changes in chemical shift and line width, which allow for the identification of the involved residues. Furthermore, analysis of the chemical shifts changes as a function of ligand concentration facilitates the determination of the respective dissociation constant.

Detailed binding information can also be obtained when the protein cannot be labeled with $^{13}C/^{15}N$. Under suitable conditions (weak binding of a small ligand to a large receptor), so-called transfer NOEs yield the receptor-bound conformation of a ligand even though the receptor is too large for NMR spectroscopy. Additionally, one can detect an interaction between a ligand and a receptor through saturation transfer, which results in intensity changes for the NMR peaks of the ligand. While the ligand is in large excess, the receptor is essentially invisible due to its dilution by a few orders of magnitude. This technique is particularly useful to screen ligands for certain target receptors. Additionally, NMR spectroscopy allows for the determination of the translational diffusion, which depends on the size and shape of the molecule. Therefore, one can determine if a protein is a monomer, a dimer, or if it forms a complex with other proteins.

Protein Folding and Misfolding NMR spectroscopic techniques also allow atomic level insight into the folding and misfolding of proteins. In combination with sub-zero temperatures (down to −15 °C) or high pressures (up to 2 kbar) it is possible to determine the 3D structure of partially folded equilibrium intermediates, and to analyze the kinetics of protein folding pathways. In addition, through combinations of H/D exchange and other NMR methods one can follow the formation of hydrogen bonds within a stable secondary structure during the folding process. Because H/D exchange and folding are competitive reactions, one can determine the velocity of the secondary structure formation from the exchange rate of individual protons. Alternatively, a so-called quenched flow apparatus permits H/D exchange reactions only during certain time periods of the folding process. Thus, a time-resolved picture of the formation of secondary and tertiary structures becomes accessible. Furthermore, ^{15}N-relaxation dispersion as well as real-time NMR methods offer unique insight into the processes of folding and misfolding of proteins.

Intrinsically Disordered Proteins Intrinsically disordered (also called natively unstructured) proteins (IDPs) lack a well-defined tertiary structure, but still fulfill critical biological functions as part of signal transduction cascades, in spliceosomes, or in cancer-associated processes. The degree of disorder ranges from completely unfolded proteins to mostly folded proteins, which contain disordered regions of 30 residues or more. IDPs possess a very distinct amino acid composition with a high abundance of proline, polar, and charged amino acids (serine, glycine, glutamate, arginine). At the same time, they are depleted in hydrophobic and aromatic amino acids that usually form the hydrophobic core of globular

proteins. Due to their dynamic nature, only NMR is capable of providing structural information about IDPs with atomic resolution.

The low sequence complexity and the absence of stable structure result in similar chemical environments for each residue. Therefore, IDPs display a narrow chemical shift dispersion (especially the amide protons) that results in severe signal overlap. Chemical shift degeneracy of C^{α} and C^{β} further limits the application of standard triple-resonance experiments (for each amino acid type the chemical shifts are close to the random coil shift) (Section 18.1.4, Triple-Resonance Experiments). Because the ^{15}N (and $^{13}C'$) dimension provide the highest dispersion in IDPs, sequential connectivities are established with 3D experiments that correlate to two nitrogen spins (e.g., HNN or (H)CANNH). Alternative approaches rely on carbon detection, which correlate the C′ spins with the directly bound nitrogen. For IDPs, the so-called NCO experiments provide improved resolution when compared to ^{15}N-HSQCs, and also allow for the detection of proline residues. Due to the lower gyromagnetic ratio, however, carbon-detected experiments are significantly less sensitive than those that use proton detection. The disordered state imparts IDPs with favorable relaxation properties. Therefore, IDPs are amenable for high dimensional 5D–7D experiments, which provide optimal resolution. In addition, the longer transverse relaxation times of IDPs enable the characterization of proteins significantly larger than 30 kDa. Even without deuteration it was possible to sequentially assign the two microtubule-associated proteins tau and MAP2c, both of which are larger than 45 kDa.

IDPs exist as ensembles of rapidly interconverting structures and therefore only provide a few NOE distance restraints (often only short-range). The secondary chemical shifts (Section 18.1.6, Determination of the Secondary Structure) obtained from the sequential assignment allow for the identification of transient secondary structural elements. Often those elements become stabilized upon binding to protein interaction partners. Long-range distance restraints to describe the structural ensemble can be obtained from RDCs and PREs (paramagnetic relaxation enhancements). To measure PREs, a paramagnetic spin label (e.g., a nitroxide) is attached to the IDP. The spin label enhances the relaxation of nearby residues (ca. 25 Å) resulting in line broadening and thus in an intensity decrease. Normalizing the reduced intensity by the intensity of the respective residue in the absence of the spin label allows for the calculation of the distance between the nucleus and the spin label, which is of course an ensemble average. Sophisticated computer programs can utilize this entire structural information together with data from small-angle X-ray scattering to calculate ensemble structures.

Structure and Dynamics of High Molecular Weight Systems and Membrane Proteins To extend the application of heteronuclear experiments to even larger proteins or protein complexes of several hundred kilodaltons, Kurt Wüthrich and colleagues developed a technique called TROSY (*t*ransverse *r*elaxation *o*ptimized *s*pectroscop*y*) at the end of the last century. The TROSY technique achieves the compensation of two relaxation mechanisms (the dipolar interaction and chemical shift anisotropy). Especially for large proteins, the resulting narrow line widths reduce the signal overlap and improve the sensitivity of the experiments. Due to its modularity, TROSY can be combined with standard 3D and triple-resonance experiments (e.g., NOESY-HSQC or HNCA) to enable a conventional sequential assignment. With this approach the structure of the 81 kDa protein malate synthase G was determined. In addition, the structure of α-helical membrane proteins, which are solubilized in detergent micelles, bicelles, or nanodiscs, can be determined using TROSY techniques. This facilitates pharmacological studies providing detailed insight into the interaction of drugs with membrane receptors.

Large protein complexes give rise to vast numbers of peaks with a great potential for chemical shift degeneracies. Therefore, to minimize spectral crowding, TROSY experiments are restricted to the analysis of an isotopically labeled subunit in an otherwise unlabeled complex, or to complexes consisting of symmetrical subunits. To circumvent the problems associated with the multitude of signals in large systems, Lewis Kay and colleagues developed the methyl-TROSY approach. This method combines the advantages of the sharp lines from the TROSY with the limited number of signals provided by selective amino acid labeling (Section 18.1.5, Selective Amino Acid Labeling). Hence, TROSY-type relaxation experiments on protonated methyl groups (isoleucine, leucine, valine, alanine, methionine, or threonine) in an otherwise perdeuterated protein allowed for the functional analysis of the gating mechanism in the proteasome.

For very large proteins (>200 kDa) the magnetization transfer through scalar coupling (also used in the TROSY-HSQC) becomes ineffective. For those proteins, the CRIPT and CRINEPT

techniques achieve an efficient magnetization transfer via cross-relaxation. Solid-state NMR represents an alternative to solution NMR methods for the study of large molecular weight systems. As the name implies, proteins are not analyzed in solution but as solid powders or microcrystals. "Magic angle" spinning of the solid protein in specially designed rotors at several thousand Hertz results in line narrowing of the resonances. Theoretically no size limitation exists; however, the complexity of the spectra for large protein restricts their analysis. Solid-state NMR mainly relies on the detection of ^{13}C and ^{15}N spins; efforts are underway to directly detect protons at very high spinning speeds ($>= 60$ kHz). Due to the low sensitivity of the respective nuclei, solid-state experiments are often restricted to two dimensions. However, the development of dynamic nuclear polarization (DNP) techniques to enhance the sensitivity holds the promise of the development of higher-dimensional experiments.

In-cell NMR Spectroscopy For structural characterization proteins are highly purified, whereas in the cell the protein coexists with other cellular parts, membranes, and thousands of other proteins at very high concentration (200–300 $g\,l^{-1}$). Therefore, it is of great interest to analyze the structure and dynamics of proteins in a cellular context. To address this need, NMR spectroscopy was applied to intact cells, so-called in-cell NMR. Initially, in-cell NMR methods were developed for bacterial cells and the structures of small proteins could be solved. More recently the focus also shifted to mammalian cells. The simplest approach for in-cell NMR is to cultivate the cells in media supplemented with isotopically labeled amino acids and to overexpress the target protein. However, metabolites often produce strong background signals and thus decrease the contrast in the resulting spectra. To obtain clear spectra, protein delivery systems were developed that introduce isotopically labeled proteins (obtained from heterologous bacterial expression) into unlabeled cells. Good delivery efficiencies might be obtained by electroporation or Lipofectamine transfection.

Currently, in-cell NMR measurements are limited to only a few hours due to acidification of the buffer and the resulting stress for the cells. Yet, this time enables the measurement of 2D HSQC spectra that can provide information about intracellular protein–protein interactions, post-translational modification, or protein dynamics. To circumvent some of the problems associated with mammalian cells, it may sometimes be more advantageous to work with cell lysates or cytoplasmic extracts. These extracts are easy to make and the reaction conditions can be better controlled, for example, when post-translational modifications are analyzed. Microinjection of proteins into large *Xenopus laevis* oocytes constitutes another alternative.

18.2 EPR Spectroscopy of Biological Systems

Olav Schiemann and Gregor Hagelüken

University of Bonn, Institute of Physical and Theoretical Chemistry, Wegelerstr. 12, 53115 Bonn, Germany

Electron paramagnetic resonance (EPR) or electron spin resonance (ESR) is a spectroscopic method that is used to obtain information on the chemical nature, structure, dynamics, and local environment of paramagnetic centers. Such centers are defined by having one or more unpaired electrons. In biological macromolecules these can be metal ions (e.g., Cu(II), Fe(III), Mn(II), Mo(V)), metal clusters (e.g., iron-sulfur clusters or manganese), or organic radicals. Organic radicals are formed, *inter alia,* as intermediates in electron transfer reactions of proteins (e.g., semiquinone anion, thiyl, or tyrosyl radicals) or they are induced by radiation damage in DNA molecules (e.g., sugar or base radicals) (Figure 18.34). Frequently, these centers are involved in catalytic cycles or in biologically relevant reactions. Diamagnetic biomolecules, in which all electrons are paired, can be made accessible to EPR spectroscopy by spin labeling techniques. In particular, nitroxides can be site specifically and covalently linked to biomolecules. This site directed spin-labeling approach together with EPR-based distance measurements between spin labels can be used to study configuration changes, to obtain coarse-grained structures of a whole biomolecule or to localize paramagnetic centers.

Similar to NMR spectroscopy, which was described in the previous section, EPR spectroscopy is a magnetic resonance technique, in which the normally degenerate $m_s = \pm ½$ levels of an electron spin $s = ½$ are split by an externally applied magnetic field, and transitions between

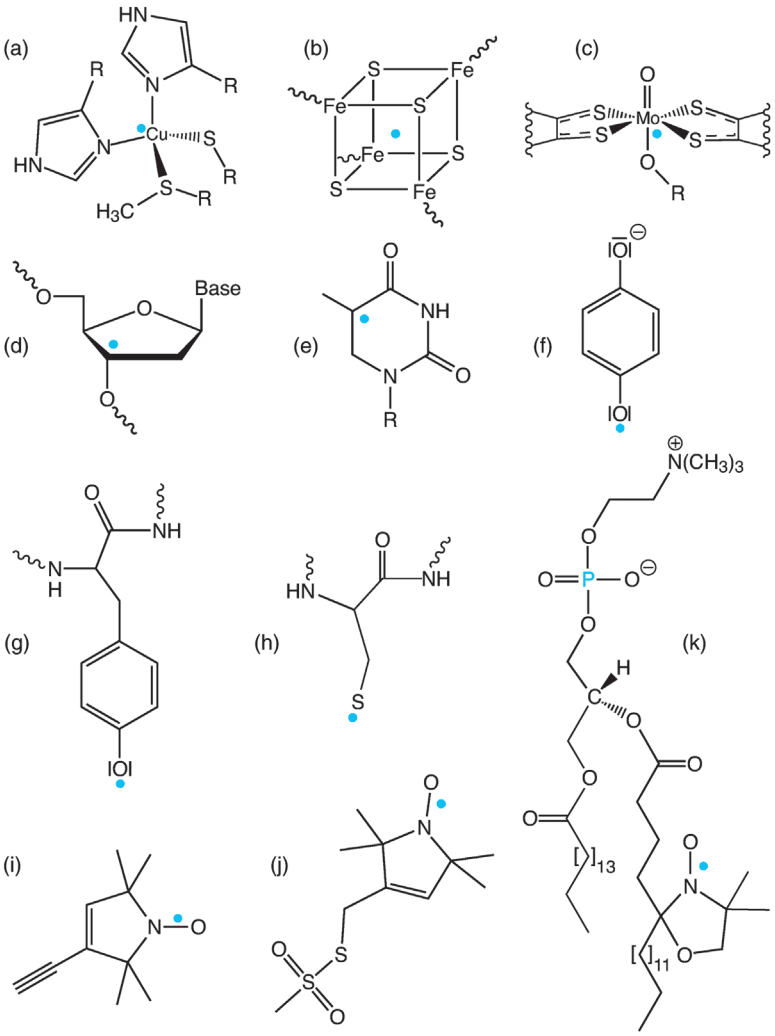


Figure 18.34 Examples of paramagnetic centers in biological systems. The wavy lines and –R represent the peptide chains, oligonucleotides, or cofactors. (a) Cu(II) in plastocyanin, (b) Fe_4S_4 clusters in ferredoxins, (c) Mo(V) in dimethyl sulfoxide reductases, (d) 3′-sugar radical in γ-irradiated DNA, (e) thymyl radical in γ-irradiated DNA, (f) benzosemiquinone anion radical, (g) tyrosyl radical in photosystem II, (h) thiyl radical in ribonucleotide reductases, (i) 1-oxyl-2,2,5,5-tetramethylpyrroline-3-acetylene (TPA), (j) 1-oxyl-2,2,5,5-tetramethylpyrroline-3-methyl) methanethiosulfonate (MTSL; nitroxide spin label for proteins), (k) 4,4-dimethyl-oxazolidine-3-oxyl-based nitroxide spin label for membranes.

these levels are induced by microwaves. The first continuous-wave (cw) EPR experiment was performed in 1944 by E.K. Zavoisky and the first pulsed EPR experiment in 1961 by W.B. Mims. The high technical requirements for pulsed EPR experiments, however, meant that pulsed EPR spectrometers only became commercially available in the late 1980s, about two decades later than for NMR. Since then, and in conjunction with the ongoing development of high frequency/high-field EPR spectrometers, computer-based EPR simulation programs, and quantum chemical methods for the translation of the EPR parameters into structural data, EPR spectroscopy has become increasingly important.

18.2.1 Basics of EPR Spectroscopy

At this point, the physical principles of EPR parameters will be briefly outlined. For a more in-depth treatment, a quantum mechanical description is inevitable. For this, the reader is referred to the references in the "further reading" section at the end of this chapter.

Electron Spin and Resonance Condition Since the Stern–Gerlach experiment it has been known that an unpaired electron has a quantum mechanical angular momentum, the so-called electron spin $\boldsymbol{s}$. The length of the vector $\boldsymbol{s}$ is given by:

$$|\boldsymbol{s}| = \hbar\sqrt{s(s+1)} \tag{18.15}$$

Here $s = ½$ is the spin quantum number and $\hbar$ is Planck's constant divided by 2π. As with any charged particle with an intrinsic angular momentum, the electron spin is linked with a magnetic moment $\boldsymbol{\mu}_\mathbf{e}$, which is oriented along the electron spin vector, but opposite to it due to its negative charge:

$$\boldsymbol{\mu}_\mathrm{e} = -g_\mathrm{e}\frac{e}{2m_\mathrm{e}}\boldsymbol{s} \tag{18.16}$$

Here, e is the elementary charge, m_e the resting mass of the electron, and $g_\mathrm{e} = 2.0023$ is the g-factor of the free electron. In an external magnetic field $\boldsymbol{B}$, there are only two possible orientations of the spins, parallel or anti-parallel to the external field. The component of the electron spin in the direction of the magnetic field B_0 (usually defined as the z-direction) is:

$$s_z = m_s\hbar \tag{18.17}$$

The magnetic quantum number m_s can take values of $+½$ and $-½$ for $s = ½$. Due to Heisenberg's uncertainty relation, the s_x and s_y components cannot be simultaneously determined with s_z. From the orientation of the electron spin along the z-axis follows a corresponding orientation of the magnetic moment $\boldsymbol{\mu}_\mathbf{e}$. Its z-component $\mu_{\mathrm{e},z}$ is given by:

$$\mu_{\mathrm{e},z} = -g_\mathrm{e}\mu_\mathrm{B}m_\mathrm{s} \tag{18.18}$$

where μ_B is the Bohr magneton, which itself is given by:

$$\mu_\mathrm{B} = \frac{e\hbar}{2m_\mathrm{e}} = 9.274\times10^{-24}\,\mathrm{J\,T^{-1}} \tag{18.19}$$

Thus, the energy E of a magnetic moment that is oriented along the z-axis of the external magnetic field is:

$$E = -\mu_{\mathrm{e},z}B_0 = g_\mathrm{e}\mu_\mathrm{B}m_\mathrm{s}B_0 \tag{18.20}$$

While, without a magnetic field, the two spin orientations $m_\mathrm{s} = -½$ and $m_\mathrm{s} = +½$ are energetically degenerate, they are split by applying an external magnetic field (Zeeman splitting). The energies E of the two orientations are given by:

$$E_{\left(m_\mathrm{s}=+\frac{1}{2}\right)} = \frac{1}{2}g_\mathrm{e}\mu_\mathrm{B}B_0 \tag{18.21}$$

$$E_{\left(m_\mathrm{s}=-\frac{1}{2}\right)} = -\frac{1}{2}g_\mathrm{e}\mu_\mathrm{B}B_0 \tag{18.22}$$

The energy difference ΔE between the two energy levels is proportional to the magnitude of B_0 (Figure 18.35):

$$\Delta E = g_\mathrm{e}\mu_\mathrm{B}B_0 \tag{18.23}$$

In a macroscopic sample with N spins, more spins are present in the low energy $m_\mathrm{s} = -½$ spin state than in the $m_\mathrm{s} = +½$ state. The ratio of the occupation numbers $N_{\left(m_\mathrm{s}=+\frac{1}{2}\right)}/N_{\left(m_\mathrm{s}=-\frac{1}{2}\right)}$ is described by the Boltzmann distribution:

$$\frac{N_{\left(m_\mathrm{s}=+\frac{1}{2}\right)}}{N_{\left(m_\mathrm{s}=-\frac{1}{2}\right)}} = \mathrm{e}^{-\frac{\Delta E}{kT}} = \mathrm{e}^{-\frac{g_\mathrm{e}\mu_\mathrm{B}B_0}{kT}} \tag{18.24}$$

For $T = 300\,\mathrm{K}$, $B_0 = 340\,\mathrm{mT}$ and $k = 1.3806\times10^{-23}\,\mathrm{J\,K^{-1}}$ (Boltzmann constant) a ratio of the occupation numbers of 0.999 can be calculated from this equation (NMR: 0.99999). Thus, at room temperature, the lower energy $m_\mathrm{s} = -½$ state is populated slightly more. Using electromagnetic radiation, spins from the $m_\mathrm{s} = -½$ level can be excited into the $m_\mathrm{s} = +½$ level, when the energy of the radiation is equal to ΔE and thereby satisfies the resonance condition:

$$h\nu = \Delta E = g_\mathrm{e}\mu_\mathrm{B}B_0 \text{ (resonance condition)} \tag{18.25}$$

Here, as in NMR spectroscopy, the magnetic field component of the electromagnetic radiation interacts with the magnetic moment of the electron spin.

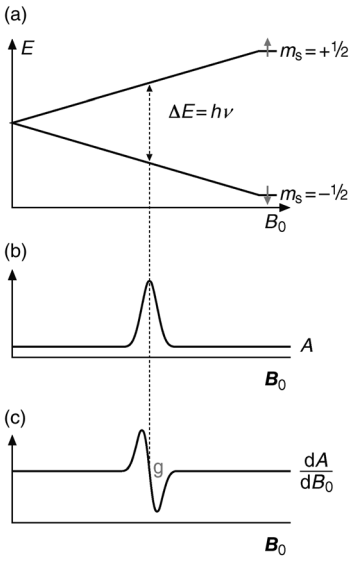


Figure 18.35 (a) Zeeman splitting for an electron spin in a magnetic field B_0. (b) The absorption line, which is obtained when the resonance condition is satisfied, and (c) the first derivative obtained by the field modulation of the absorption line.

18.2.2 cw- EPR Spectroscopy

In cw-EPR spectrometers, microwaves at a constant frequency are continuously applied to the sample and the external magnetic field is swept. The experiment is set-up such that at

some point during this process the resonance condition is met. Most commonly, cw-EPR spectrometers operating at a microwave frequency of about 9.5 GHz (X-band) are used, so that for a radical with $g \approx 2$ the resonance absorption occurs at a magnetic field of about 340 mT. Since the ratio of the occupation numbers is close to one, the absorption signal is of relatively low intensity. In cw-EPR spectrometers, the sensitivity is increased by adding a small modulated magnetic field component to the external magnetic field, which enables the lock-in detector to filter-out noise. This field modulation is also the reason that the absorption signal is obtained in the form of its first derivative (Figure 17.35). In addition, the sensitivity can be further improved by making measurements at low temperatures and stronger magnetic fields (increase of the occupation ratio). To indicate the position of the line independently of the magnetic field and the microwave frequency, the position of the absorption line is expressed as a g-value (similar to the chemical shift of NMR spectroscopy):

$$g = \frac{h\nu}{\mu_B B_0} = \left(7.144775 \times 10^{-2}\right) \frac{\nu}{B_0} (\text{in MHz/mT}) \qquad (18.26)$$

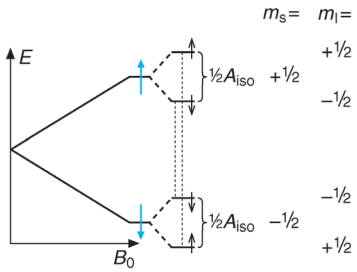


Figure 18.36 Splitting scheme for an unpaired electron (blue), which is strongly coupled to a ^{1}H nucleus ($I = ½$, black); A_{iso} is assumed to be positive. The dashed lines show the transitions, which are allowed by the EPR selection rules ($\Delta m_s = \pm 1$ and $\Delta m_I = 0$). Thus, in the EPR spectrum two lines, centered around g_{iso} and separated by A_{iso}, would be observed.

18.2.3 *g-Value*

For a free electron only one line would be observed, with $g = g_e = 2.0023$. However, if the unpaired electron resides in a molecule, deviations from g_e occur due to spin–orbit coupling. These deviations are characteristic for the electronic state, the bonding situation, and the geometry of the particular molecule. For organic radicals, these deviations are usually small, since their magnetic orbital moment is nearly zero. The nitroxide TPA (Figure 18.34i), for instance, has a g-value of 2.006. For transition metal ions, however, significantly larger deviations can be observed. The Fe^+ in an MgO matrix, for example, has a g-value of 4.1304. The g-values can thus be used to characterize and distinguish different paramagnetic species. A quantitative calculation or a translation of observed g-values into structural parameters with, for example, density functional theory (DFT) methods is, however, still complicated, especially for transition metal ions, and is the subject of current research.

18.2.4 Electron Spin Nuclear Spin Coupling (Hyperfine Coupling)

In addition to the spin–orbit coupling, the magnetic moment of the electron spin can be coupled to the magnetic moments of nuclei (Figure 18.36) if the nuclear spin I is greater than zero (e.g., ^{1}H, ^{14}N, ^{31}P, ^{13}C, ^{17}O, ^{33}S, ^{55}Mn, ^{95}Mo, ^{57}Fe, ^{51}V). The coupling of the electron and the nuclean magnetic moments leads to a splitting of the absorption line into $M = 2NI + 1$ lines (multiplicity rule). Here, N is the number of equivalent magnetic nuclei and I is their nuclear spin quantum number.

The magnitude of the splitting, which is called the hyperfine coupling constant A_{iso}, depends linearly on the magnetic moment μ_I of the coupling nucleus and the spin density $|\psi(r = 0)|^2$ of the unpaired electron at the nucleus (Fermi contact interaction):

$$A_{iso} \sim M_I |\psi(r = 0)|^2 \qquad (18.27)$$

Since only s-orbitals have a non-zero value at the nucleus but the unpaired electron in most radicals resides in a p-, π, or d orbital, the question arises as to why such radicals show a hyperfine coupling in the first place. The reason for this is spin polarization, a mechanism that generates spin density in low energy s orbitals (Figure 18.37).

If an EPR spectrum of a radical is obtained with resolved hyperfine coupling structure, the spin density distribution can be determined from the hyperfine coupling constants and thus statements regarding the nature and structure of the radical can be made. An example is the cw-EPR spectrum of the nitroxide TPA (Figure 18.38a). As in all alkyl nitroxides the unpaired electron resides in a π orbital between the nitrogen and oxygen atom. At both nuclei, spin density is generated through spin polarization. However, since the most abundant oxygen isotope, ^{16}O, has a nuclear spin of $I = 0$, no line splitting is induced by this nucleus. In contrast, nature's most abundant nitrogen isotope, ^{14}N, has a nuclear spin of $I = 1$, so the absorption line is split into a triplet with a hyperfine coupling constant

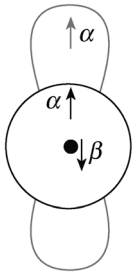


Figure 18.37 Mechanism of spin polarization: the unpaired electron with an α-spin in the p-orbital has its largest probability at a relatively large distance from the nucleus. According to Hund's rule, an electron in the s orbital has an electron spin (α) parallel to the unpaired electron in the p-orbital, if it is also far away from the nucleus. Due to Pauli's rule, the second s-electron close to the nucleus must then be set anti-parallel to the first s-electron (β). In this way, excess β-spin density is induced at the nucleus, since the probability for α- and β-s electrons at the nucleus is no longer equal.

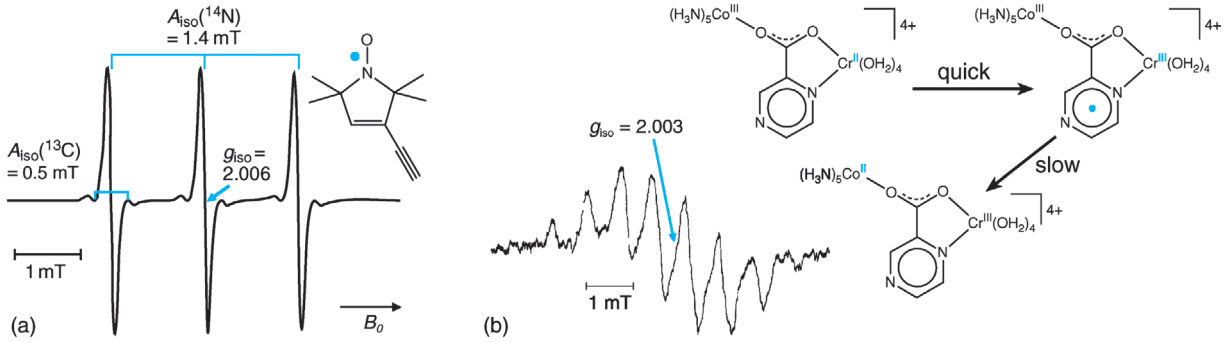


Figure 18.38 (a) cw-X-band EPR spectrum of TPA in liquid solution. (b) Model for an electron transfer chain. The reduction of Co(III) to Co(II) is carried out by an intramolecular electron transfer. The electron is first transferred from the Cr(II) to the pyrazine and only then to the Co(III). The pyrazine radical could be detected by EPR spectroscopy on the basis of the hyperfine signature and the *g*-value. Often the cw-EPR spectra are shown without an *x*-axis and only a scale ruler is given. Source: Spiecker, H. and Wieghardt, K. (1977) *Inorg. Chem.*, **16**, 1290–1294. With permission, Copyright © 1977, American Chemical Society.

$A_{iso} = 1.4\,mT = 39.2\,MHz$. The additionally observed small lines on the low and high field side of each ^{14}N-line are caused by ^{13}C hyperfine coupling to one of the directly bonded carbon nuclei ($I = ½$) or one of the carbon nuclei of the methyl groups. This hyperfine coupling of 0.6 mT splits each of the three lines into a doublet. However, due to the low natural abundance of ^{13}C (1.1%), the intensity of this triplet of doublets is low, so that the ^{14}N triplet dominates the EPR spectrum. Actually, each of the three ^{14}N lines would have to be split by six ^{13}C nuclei. However, the probability that multiple ^{13}C isotopes are found within one molecule is so small that these couplings are not observed.

Although, the question of the spin density distribution may appear merely academic for a nitroxide, it is important for cofactors in electron transfer proteins. For example, to understand electron transfer mechanisms in proteins, it is crucial to know whether the transported electron can reside on a cofactor in an electron transfer pathway or which part of a co-factor acts as the electron acceptor or electron donor (Figure 18.38b).

18.2.5 *g* and Hyperfine Anisotropy

The considerations in the preceding sections are based on radicals in liquid solution, that is, the radicals rotate so fast compared to the EPR time scale (actually compared to the size of the anisotropy of the interaction) that they do not a have fixed preferential orientation with respect to the external magnetic field B_0. Spectra acquired under such conditions are called isotropic. If the sample is frozen, a powder or a single crystal, each molecule has a fixed orientation with respect to B_0 and the EPR spectra are characterized by orientation-dependent (anisotropic) contributions. Such contributions can be found for *g*, the hyperfine coupling and the coupling between the unpaired electrons (Section 18.2.6). Although the occurrence of these anisotropic contributions in an EPR spectrum renders its analysis more difficult, they also offer more detailed information about the nature and structure (angles and distances) of the paramagnetic system. For this reason, as well as to increase the sensitivity and to extend the lifetime of short-lived radicals; most EPR experiments on biological systems are performed in frozen solution at very low temperatures (down to, for example, 3.5 K).

***g* Anisotropy** If the unpaired electron resides in an s orbital, for which, due to its spherical symmetry, the three spatial directions *x, y,* and *z* are equivalent, then $g_x = g_y = g_z$. For EPR spectra in such a spherically symmetric case and in the absence of hyperfine interactions, one observes one line with a *g*-value at g_{iso} even in the solid state. Anisotropy in *g* occurs when the unpaired electron resides in an orbital of the molecule that is not spherical. In the axisymmetric case (two equivalent directions in space), a spectrum as shown in Figure 18.39a with two canonical *g*-values ($g_\perp$ and $g_\parallel$) is obtained. In the orthorhombic case ($x \neq y \neq z$) all three *g*-values are different ($g_x \neq g_y \neq g_z$) and one obtains a spectrum similar to the one shown in

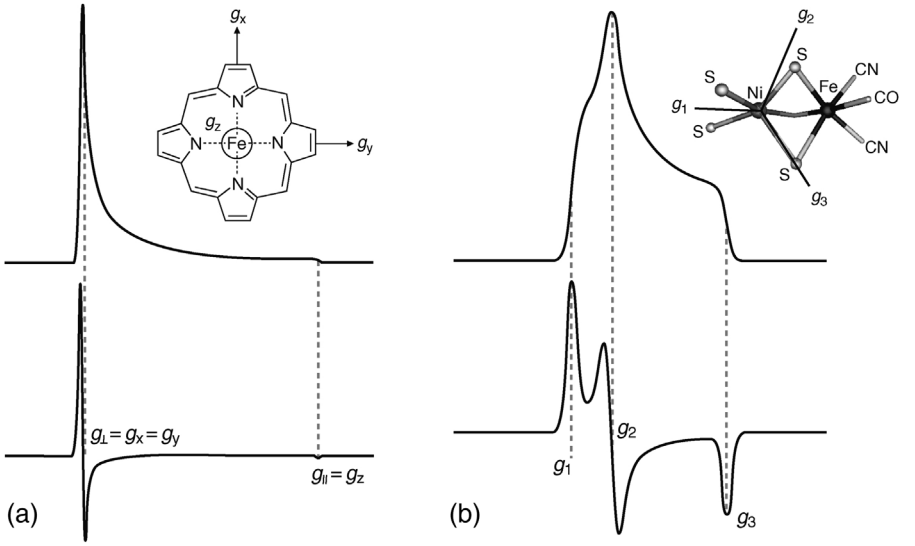


Figure 18.39 (a) Axial cw-X-band EPR spectrum. The first derivative is given at the bottom and, for clarity, the absorption spectrum at the top. The absorbance at $g = g_z = g_\parallel$ results from molecules that are oriented with $g_\parallel$ parallel to B_0, while at $g_x = g_y = g_\perp$ only those molecules absorb that are oriented with $g_\perp$ parallel to B_0. Between these two points, only those molecules contribute to the absorption that have neither $g_\parallel$ nor $g_\perp$ parallel to B_0. An example of an axially symmetric molecule with resolved $g_\parallel = 2.00$ and $g_\perp = 5.67$ is the high-spin Fe(III) porphyrin of cytochrome P450. Source: from Woggon, W.-D. *et al.* (1998) *Angew. Chem.*, **110**, 3191–3195. With permission, Copyright © 1998 WILEY-VCH Verlag GmbH, Weinheim, Germany. (b) Orthorhombic cw-X-band EPR spectra. A system with orthorhombic symmetry and g-values at $g_1 = 2.196$, $g_2 = 2.145$, and $g_3 = 2.010$ obtained from the Ni center in the [Ni -Fe] hydrogenase. Source: Foerster, S. *et al.* (2005) *J. Biol. Inorg. Chem.*, **10**, 51–62. With permission, Copyright © 2004, SBIC. The assignment of *x, y,* and *z* to the corresponding *g*-values is possible by measurements on single crystals.

Figure 18.39b. The indices *x, y*, and *z* stand for the three canonical directions in space, meaning that g changes in accordance with the orientation of the molecule with respect to B_0. In liquid solution, however, when the molecule rotates rapidly, the anisotropic components of g are averaged out such that:

$$g_{iso} = \frac{1}{3}\left(g_x + g_y + g_z\right) \tag{18.28}$$

or:

$$g_{iso} = \frac{1}{3}\left(2g_\perp + g_\parallel\right) \tag{18.29}$$

High Field/High-Frequency EPR Spectrometer In many cases, especially in organic radicals, the three g-values are not separated at a magnetic field of 340 mT/microwave frequency of 9.5 GHz (X-band) due to the relatively small g-anisotropy. But it is possible to resolve the g-anisotropy by using higher frequencies/fields (Figure 18.40). This exploits the fact that the separation of the g-values (measured in magnetic field units) increases with the magnetic field. In contrast, the size of the line splitting due to the hyperfine coupling is independent of the magnetic field. EPR spectrometers working at higher frequencies and that are commercially available are those working at 36 GHz (Q-band)/1.3 T, 95 GHz (W-band)/

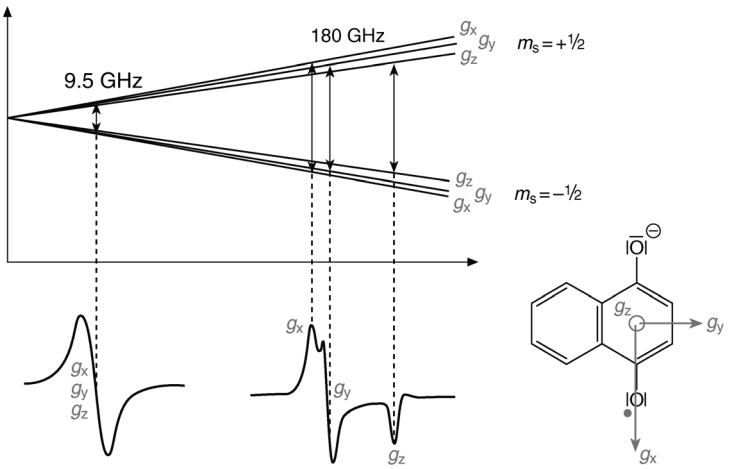


Figure 18.40 Magnetic field dependent splitting of g_x, g_y, and g_z. An example of the cw-EPR spectra of a semiquinone anion radical in X(9.5 GHz/0.34T)- and G (180 GHz/T 6,4)-band. The g-values themselves do not change, only the distance (in mT) between them increases.

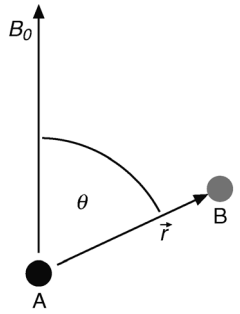


Figure 18.41 Illustration of the distance vector ***r*** and the angle θ. A and B can be either an electron and a nucleus or two electrons (Section 18.2.6).

3.4 T, or 260 GHz/9.2 T. Spectrometers with even higher frequencies (360 GHz, 640 GHz, and in the THz range) are also used, but are technically very demanding and not yet commercially available. However, the technical effort is worthwhile, not only for the resolution of the *g*-values, but also because of the following advantages:

- Superimposed spectra of different radicals can be separated by their different *g*-values.
- The sensitivity of the spectrometer increases at higher magnetic field/frequency (due to the larger population difference), which means that smaller sample amounts are needed (e.g., at 180 GHz, 0.1 µl of a 1 mM Mn(II) solution is sufficient).
- If the same sample is measured at different microwave frequencies, the magnetic-field-independent hyperfine splitting constant can be separated from the magnetic field dependent *g*-splitting (in magnetic field units).

Hyperfine Anisotropy An anisotropic hyperfine coupling component A_i, which is observed in a spectrum, is the sum of the isotropic hyperfine coupling constant A_{iso} and an anisotropic, purely dipolar component $A_{i,dip}$. The subscript i stands for one of the three spatial directions x, y, or z, and means that the hyperfine coupling, depending on the spatial direction or orientation of the radical, varies relative to B_0:

$$A_i = \left(A_{iso} + A_{i,Dip}\right) \tag{18.30}$$

The anisotropic portion of the hyperfine coupling arises from the dipole–dipole coupling between the magnetic moments of the electron and the nucleus. It depends both on the distance r between the electron and the nucleus as well as on the angle θ between the distance vector ***r*** and the external magnetic field B_0 (Figure 18.41).

Depending on the symmetry of the paramagnetic center, a distinction of three special cases for the hyperfine coupling is made: spherical symmetry with $A_x = A_y = A_z$, axial with $A_\perp$ (= $A_x = A_y$) and $A_|$ (= A_z), and orthorhombic with $A_x \neq A_y \neq A_z$ (Figure 18.42). In liquid solution the dipolar hyperfine coupling components cancel each other out, so that:

$$A_{iso} = \frac{1}{3}\left(A_x + A_y + A_z\right) \tag{18.31}$$

$$A_{iso} = \frac{1}{3}\left(2A_\perp + A_\parallel\right) \tag{18.32}$$

18.2.6 Electron Spin–Electron Spin Coupling

If, within a biomolecule, two unpaired electrons A and B are located at a fixed distance r to each other, a coupling ν_{AB} between these two electrons can occur. As seen for the hyperfine coupling constant, this electron–electron coupling is the sum of an isotropic and an anisotropic component:

$$\nu_{AB} = J + D \tag{18.33}$$

The isotropic part, J, is the exchange interaction, which is non-zero when the wave functions of both electrons overlap ($R < 10$ Å) or if the two electrons are interacting via a conjugated bridge

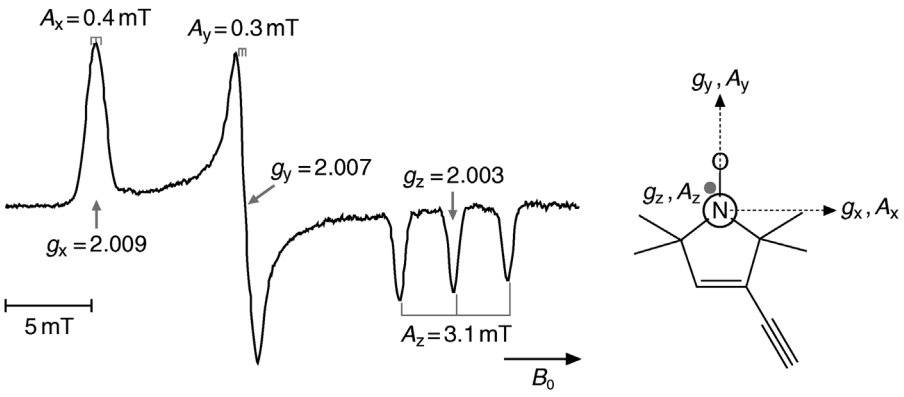


Figure 18.42 Hyperfine and *g* anisotropy for the example of the cw-G (180 GHz/6.4 T)-band EPR spectrum of TPA in frozen solution. The A_x and A_y hyperfine couplings are not resolved.

(super exchange). It can be observed in liquid solution and allows to draw conclusions as to whether the two electron spins are aligned parallel (ferromagnetic) or antiparallel (antiferromagnetic) to each other. The anisotropic component D of the coupling is based on the interaction between the magnetic dipoles of the two electrons and is orientation-dependent. It can only be observed when the biomolecules have a fixed orientation with respect to B_0 (frozen solution, powder, or single crystal). In liquid solution, the anisotropic component is averaged out:

$$D = \nu_{\text{Dip}}\left(1 - 3\cos^2\theta\right) \tag{18.34}$$

$$\nu_{\text{Dip}} = \frac{m_B^2 g_A g_B \mu_0}{4\pi h} \times \frac{1}{r_{AB}^3} \tag{18.35}$$

where: θ is the angle between the distance vector $\boldsymbol{r}$ and the external magnetic field B_0 (Figure 18.41);
g_A and g_B are the g-values of the two electrons;
μ_0 is the permeability in vacuum;
m_B is the Bohr magneton.

The first term in Equation 18.35 (ν_{Dip}) is therefore a constant.

In an anisotropic cw-EPR spectra, the electron–electron dipolar coupling manifest itself as an additional line splitting if it is larger than the line width. If the dipolar coupling constant ν dip can be determined from a spectrum, then the distance between the two unpaired electrons can be calculated. Using cw-EPR spectra, distances of up to 20 Å can be determined in this way. At greater distances, the dipolar coupling is obscured by the intrinsic linewidth and pulsed methods can be used (see also Section 18.2.7). Such distance measurements are important in the determination of the arrangement of paramagnetic centers in biological macromolecules (Figure 18.43). Similarly, the distance between two subunits of a protein can be measured by site-directed spin labeling of the subunits with two nitroxides and concomitant determination of the dipolar coupling between the two spin labels. By varying the spin label positions, and measuring the distances between the respective pairs, it is possible to figure out the spatial arrangement of the subunits. In the same way one can examine whether the binding of ligands (small organic molecules, metal, proteins, RNA, DNA) leads to global structural changes.

18.2.7 Pulsed EPR Experiments

With cw-EPR methods, strongly coupled nuclei in the vicinity of the unpaired electron can be characterized. Pulsed EPR methods allow the resolution of weaker couplings to more distant nuclei or distant unpaired electrons, which in the cw spectrum are hidden underneath the broad spectral lines. Other advantages of pulsed experiments are that the pulse sequences allow selection of those contributions to the spectrum that one would like to analyze and that multidimensional experiments are possible. In particular, with experiments such as ESEEM (electron spin echo envelope modulation) and pulse ENDOR

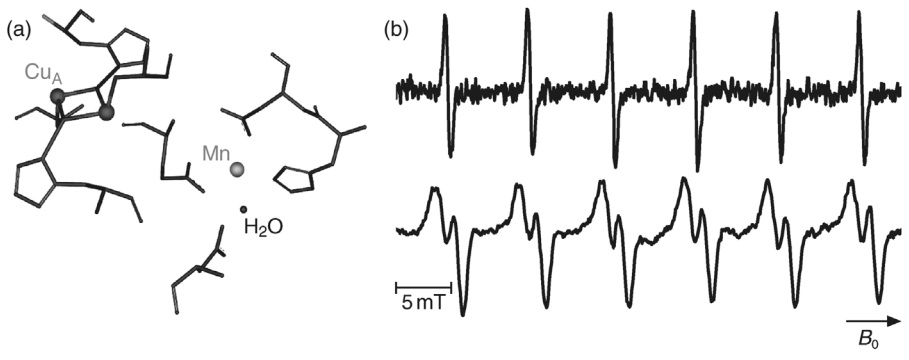


Figure 18.43 Example of a biological system with electron–electron dipolar coupling. (a) Arrangement of the binuclear Cu_A center and the Mn(II) ion in cytochrome c oxidase from *Paracoccus denitrificans*. (b) cw-W-band (94 GHz/3.3 T) EPR spectra of the Mn(II) center. The Mn(II) ion ($s = {}^5/_2$ and $I = {}^5/_2$) leads to a W-band cw-EPR spectrum with six lines. In this case, the Cu_A center has been switched into the diamagnetic $s = 0$ state, so that it cannot be detected by EPR spectroscopy. If the sample conditions are then adjusted such that the Cu_A center turns into the $s = ½$ state, a cw-EPR spectrum can be observed in which each manganese line is split into a doublet. From the magnitude of the splitting (2.2 mT) the dipolar coupling constant was determined to be 3.36 mT and with this a Cu_A–Mn distance of 9.4 Å could be calculated. The paramagnetic Cu_A center itself was not detected directly. Source: Käß, F. *et al*. (2000) *J. Phys. Chem. B*, **104**, 5362–5371. With permission, Copyright © 2000, American Chemical Society.

(electron nuclear double resonance) detailed conclusions about the structural environment of the unpaired electron can be made (within a radius of up to 10 Å). With PELDOR (pulsed electron–electron double resonance), even distances of up to 8 nm between two electron spins can be measured.

Larmor frequency: Frequency with which an electron or nucleus spins about the z-axis, the direction of an externally applied magnetic field. The name commemorates the Irish physicist Sir Joseph Larmor.

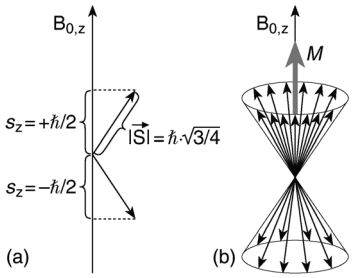


Figure 18.44 (a) Illustration of the spin orientation relative to B_0. (b) Illustration of the magnetization M (big arrow). The black arrows represent the magnetic moments $\mu_{e,z}$ of the electron spins.

Basics The basics of pulsed EPR spectroscopy are comparable to those of pulsed NMR spectroscopy (Section 18.2.1). For EPR, these considerations must simply be transferred to the electron spin.

In Section 18.2.1 it was shown that, for a given electron spin, the length of the s-vector and its z-component are defined, while the x and y components are undefined. Note that the magnetic moment of the spins is thus not exactly parallel to B_0 (Figure 18.44a) – even if we say so in the following. However, the presence of the magnetic field leads to the electron spins precessing on cones parallel and anti-parallel to the B_0 magnetic field axis (Figure 18.44b). The magnetic-field dependent frequency of this precession is known as the Larmor frequency. Since, according to the Boltzmann distribution, more spins are oriented with their magnetic moments μ_e parallel to B_0, the sample experiences a macroscopic magnetization M parallel to B_0.

Pulses To generate microwave pulses, the microwave radiation is turned on, and quickly turned off again, after a short period of time. This corresponds to a rectangular microwave pulse. In EPR spectroscopy, such pulses typically have pulse lengths in the range of nanoseconds (NMR: microseconds). Because of the very short pulse lengths, the pulses do not have a singular microwave frequency (such as the microwaves in the cw experiment), but a certain range of frequencies (Heisenberg uncertainty principle). It follows that, at a constant magnetic field, a larger part of the EPR spectrum can be excited. In the following, 90°- or 180°-pulses designate such pulses, which rotate the electron spin by 90° or 180°. In this case, all pulses are irradiated along the x-axis (from $-x$ to $+x$) and the magnetization is rotated clockwise. Both pulse width and pulse amplitude affect the magnitude of the rotation angle. Thus, a 180° pulse has either twice the length or a twofold power of a 90° pulse.

If a 90° pulse is applied to a sample, then this pulse interacts with the spins and rotates M from the z to the $+y$-axis (Figure 18.45).

In Section 18.2.1 it was stated that spins can only align parallel or antiparallel to the field. This raises the question of how the individual spins must be aligned to cause a magnetization in the $+y$ direction. This can be illustrated in the following way: The 90° pulse induces an equal population of spins on both energy levels, making the individual magnetic moments along the z-axis add up to zero. Simultaneously, the spins are no longer uniformly distributed on either of the two precession cones but form spinning packages in the direction of the $+y$ axis. Since the spins in these packages have the same orientation and speed, this phenomenon is called phase coherence (Section 18.1.1 and Figure 18.5). The spinning packages are not static along the $+y$-axis, but precess at their Larmor frequency around the z-axis. In contrast to the cw experiment in which the absorption of the microwave is detected, this precession of the spins induces a current or a signal in the detector, which can then be recorded.

Relaxation In general, relaxation means the return from the excited state to the ground state. Due to random spin–spin interactions, the phase coherence described above is lost with time, so that the spins are again evenly distributed in the two precession cones. This loss of phase coherence is called T_2 relaxation. For electron spins, T_2 is in the range of nano- to microseconds. At the same time, due to energy exchange processes with the environment, the spins return again into the Boltzmann equilibrium distribution between the two energy states. This relaxation is called T_1 relaxation and for electron spins it lies in the range of micro- to milliseconds.

Spin Echoes For technical reasons, and for most samples, no free induction decay (FID, Section 18.1.2) can be observed after the 90° pulse in EPR spectroscopy. Instead, as a

workaround, a so called "spin echo" is detected. Below, a simple two-pulse sequence is described that generates such a spin echo (Figure 18.46).

At a time τ after the 90° pulse, the spin packets have traveled different distances from the *y*-axis due to their differing rotational velocities (Figure 18.45c). The different rotation speeds are the result of the slightly different magnetic environments and thus Larmor frequencies of the spins. If the sample is now irradiated with a 180° pulse, the spins (Figure 18.45d) are rotated by 180° around the *x*-axis. Since they maintain their direction and speed of rotation, the spins start to re-phase (Figure 18.45e) and form a magnetization or a spin echo along the *y*-axis after the time 2τ (Figure 18.45f). In honor of its discoverer Erwin Hahn, this echo is called Hahn echo.

A different spin echo can be generated with a three-pulse sequence (Figure 18.47). With a 90° pulse, the magnetization is rotated to the *x/y* plane, then after a short time τ a second 90° pulse follows. By this second 90° pulse, the *x/y* magnetization is stored along the *z*-axis. With a third 90° pulse after a time interval *T,* the magnetization is detected as a so-called stimulated echo.

ESEEM – Electron Spin Echo Envelope Modulation If the time interval τ of the pulse sequence shown in Figure 18.46 is gradually increased, and the amplitude of the Hahn-echo is detected at each step, one can observe an oscillating amplitude of the echo, which is known as electron spin echo envelope modulation (ESEEM). The cause of this modulation is transitions between the nuclear spin levels. A plot of the echo amplitude versus τ yields a time trace from which the corresponding oscillation frequencies can be obtained by Fourier transformation. From these frequencies, weak couplings of nuclei, which are too small to be observed by cw-EPR, can be determined and thus more detailed information about the structure of the environment of the paramagnetic center can be obtained. A disadvantage of this pulse sequence is that due to the rapid T_2 relaxation the line width is very large, resulting in overlapping lines. In addition, higher harmonic frequencies occur, and the observed redundant sum and difference frequencies make it difficult to interpret the spectra.

These problems can be avoided by using the stimulated echo based three-pulse ESEEM. In this experiment, the time interval T in Figure 18.47 is gradually increased and the echo amplitude of the stimulated echo is recorded as a function of T. Because the amplitude of the stimulated echo decays with the slower relaxation time T_1, a longer time window for monitoring the oscillation is available and thus the obtained frequency lines are narrower than with the two-pulse ESEEM.

An example of a three-pulse-ESEEM is an experiment on the electron transfer protein bo_3 ubiquinol oxidase from *Escherichia coli*. During the electron transfer reaction, the ubiquinone is converted into an ubisemiquinone anion radical ($UB^{\bullet -}$), which enabled an EPR investigation into how the $UB^{\bullet -}$ is structurally bound. In Figure 18.48 the corresponding three-pulse ESEEM is shown in both the time and the frequency domains. In the Fourier transformed spectrum, there are four lines at 0.95, 2.32, 3.27, and 5.2 MHz, which were assigned to a ^{14}N-nucleus near the $UB^{\bullet -}$.

The occurrence of these four lines can be explained as follows: Since ^{14}N has a nuclear spin of $I = 1$, each of the two electron spin states $m_s = \pm ½$ split into three levels due to the ^{14}N hyperfine coupling (Figure 18.49). If the nuclear Zeeman interaction (interaction between the magnetic moment μ_I of the nuclear spin and B_0) is half as large as the hyperfine coupling, then both interactions cancel out in one of the two m_S levels. The energy separation in this m_S level is then determined by the quadrupole interaction that occurs for nuclei with $I > ½$. This quadrupole interaction is defined by the quadrupole coupling constant Q and the asymmetry parameter η. Both parameters are very sensitive to the distribution of charge in a molecule and the molecular structure. Between the three nuclear spin levels in each of the two m_s-levels there are three nuclear spin transitions. The three nuclear spin transitions ν_0, ν_+, and ν_- are particularly intense in three-pulse ESEEM. The double-quantum transition ν_{dq} from the other m_S-level, however, is of low intensity and broad and the two single quantum transitions ν_{sq1} and ν_{sq2} are often not detected.

From the frequencies belonging to the four lines, the isotropic ^{14}N hyperfine coupling constant, the ^{14}N quadrupole coupling constant, and the asymmetry parameter η could be calculated. With these data and from the two-dimensional spectrum (Figure 18.51b below) and

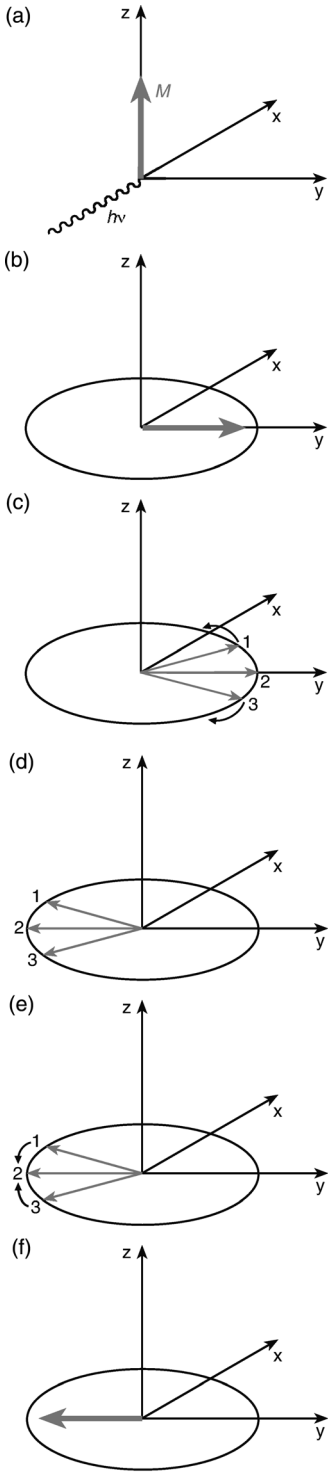


Figure 18.45 (a) Magnetization *M* along B_0; *x, y,* and *z* are Cartesian coordinates and B_0 is oriented along *z;* (b) *M* after the 90° pulse; (c) dephasing during the time interval τ; (d) inversion of the spins by the 180° pulse; (e) refocusing of the spins in the time interval τ following the 180° pulse; (f) Hahn-echo at time 2τ after the 90° pulse.

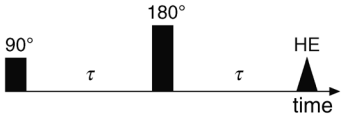


Figure 18.46 Two-pulse sequence to generate a Hahn echo (HE).

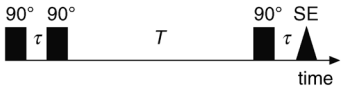


Figure 18.47 Three-pulse sequence for generating a stimulated echo (SE).

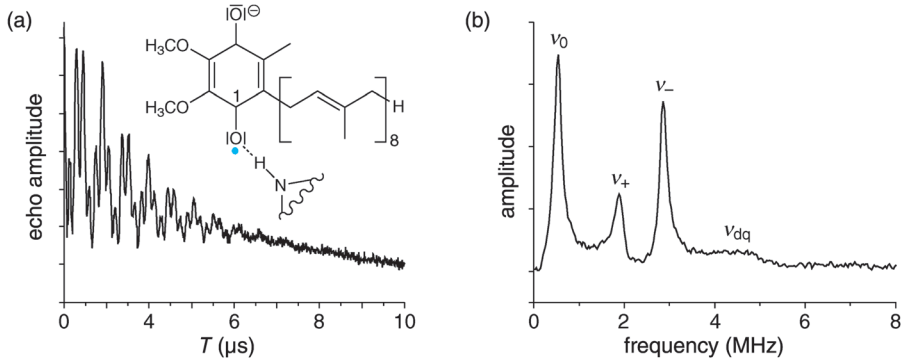


Figure 18.48 Three-pulse ESEEM spectrum and structural formula of $UB^{\bullet -}$ in the QH-binding pocket of the ubiquinol oxidase. (a) Time domain spectrum and (b) frequency domain spectrum after Fourier transform of (a). Source: reproduced with permission from Grimaldi, S. *et al.* (2001) *Biochemistry*, **40**, 1037–1043. With permission, Copyright © 2003, American Chemical Society.

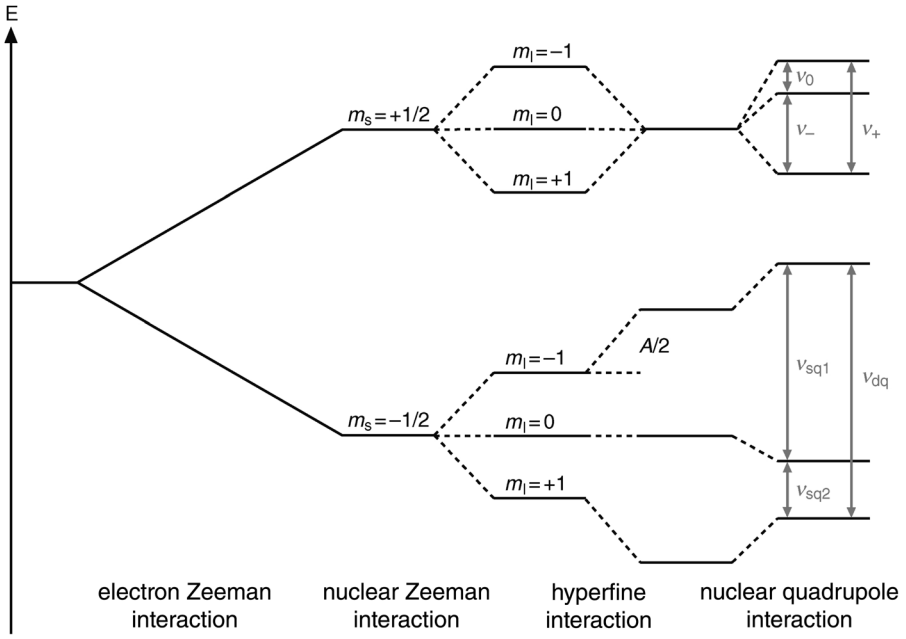


Figure 18.49 Splitting scheme for a coupled spin system with $s = ½$ and $I = 1$, for the case that the hyperfine interaction A_{iso} and the nuclear Zeeman interaction ν_N cancel each other out ($A_{iso} = 2\nu_N$).

isotope labeling it was concluded that the $UB^{\bullet -}$ is bound only with the C1 carbonyl group via a strong hydrogen bond to a nitrogen atom in the amino acid backbone while the other carbonyl group is not bound to the protein. This finding of an asymmetric binding of $UB^{\bullet -}$ helped in understanding the directional electron transfer in this protein.

HYSCORE – Hyperfine Sublevel Correlation Experiment The hyperfine sublevel correlation experiment (HYSCORE) is a two-dimensional cross-correlation experiment, which can be used to identify nuclear spin transitions (lines) belonging to the same nucleus.

The experiment is carried out such that, between the last two 90° pulses of a stimulated echo sequence, a 180° mixing pulse is introduced (Figure 18.50). Then, both the time interval t_1 and the time interval t_2 are varied. After a two-dimensional Fourier transform a two-dimensional frequency domain spectrum is obtained. In this spectrum, cross correlations occur between those peaks that belong to nuclear spin transitions in different m_s levels but from the same nucleus. In the case of ubiquinol oxidase, HYSCORE was used to determine whether the four lines in the three-pulse ESEEM (Figure 18.48) originate from the nitrogen nucleus. The

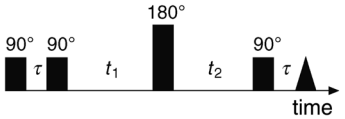


Figure 18.50 HYSCORE pulse sequence.

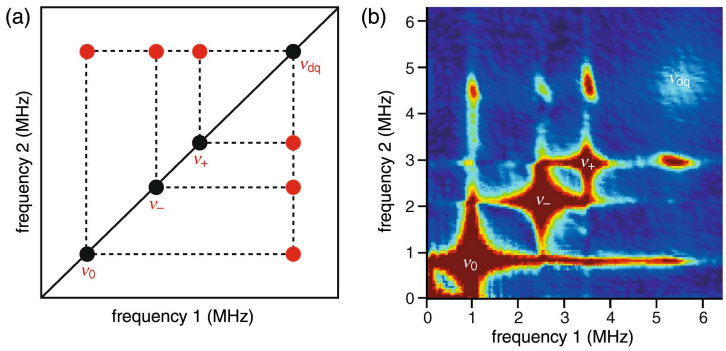


Figure 18.51 (a) Theoretical HYSCORE spectrum. (b) HYSCORE spectrum of $UB^{\bullet -}$ in the Q_H binding pocket of bo_3 ubiquinol oxidase from *Escherichia coli*. For a better overview only one of the four quadrants is shown. Source: reproduced with permission from Grimaldi, S. *et al.* (2001) *Biochemistry*, **40**, 1037–1043; © 2003, American Chemical Society.

corresponding HYSCORE spectrum is shown in Figure 18.51b. One can clearly see the cross-correlations between the four lines indicating that indeed they can be assigned to a single nitrogen nucleus.

ENDOR – Electron Nuclear Double Resonance Electron nuclear double resonance (ENDOR) experiments can be used to observe weak and strong hyperfine couplings of near and distant centers. There are several ENDOR variants that can be carried out both as cw- or pulse-experiments.

In a Davies ENDOR experiment (Figure 18.52a), the magnetization is inverted with a 180° pulse (from $+z$ to $-z$), and after a mixing time T, the magnetization is detected again with a 90°-τ-180°-τ-echo sequence. During the time T, a 180° radio frequency pulse is applied (180° pulse for the nuclear spins) and the echo amplitude is measured as a function of the radio frequency. If nuclear spin transitions are induced by the radio frequency, then lines occur at the respective radio frequencies in the ENDOR spectrum.

Another pulsed ENDOR experiment, Mims ENDOR (Figure 18.52b), is based on the stimulated echo sequence, wherein the radio frequency pulse is irradiated after the second 90° pulse. This pulse sequence has a higher sensitivity for small hyperfine couplings. Both pulse sequences are static with respect to the time axis, that is, no time interval is changed during both experiments. Only the frequency of the radio wave is changed while monitoring the amplitude of the detected echo. Figure 18.53b shows a Mims ENDOR spectrum of a ^{31}P nucleus in a phospholipid.

Amongst others, ENDOR experiments are interesting for three reasons:

1. The resolution is very good, which is why even small hyperfine couplings and hyperfine anisotropies can be resolved.
2. The number of peaks is reduced compared to the cw-EPR spectra. If N magnetically inequivalent ^1H nuclei are contributing to a cw-EPR experiment, then the number of lines is 2^N, while in the corresponding ENDOR experiments only $2N$ peaks will be present. In the case of N magnetically equivalent ^1H nuclei a cw-EPR spectrum will show $N+1$ lines, whereas an ENDOR spectrum shows only two lines.

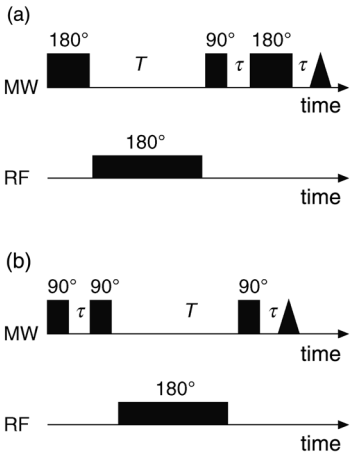


Figure 18.52 ENDOR-pulse sequence: (a) Davies sequence and (b) Mims sequence. RF stands for radio frequency and MW for microwave frequency.

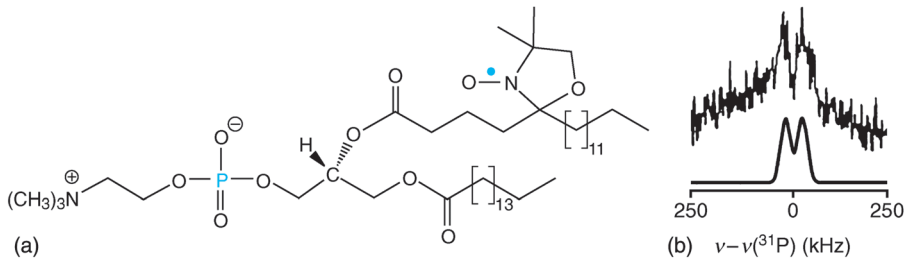


Figure 18.53 (a) Structural formula of the phospholipid and (b) Mims ENDOR spectrum of the ^{31}P nucleus and simulation of the spectrum (below); $\nu(^{31}P)$ is the free Larmor frequency of the ^{31}P-nucleus (6 MHz). From the splitting of 33 kHz and the line width, the distance between the ^{31}P-nucleus and the electron spin was determined to be 1 nm. This corresponds to an extended conformation of the lipid. The small signal-to-noise ratio indicates that the distance of 1 nm is at the upper distance limit of the method. Source: adapted from Zänker, P.P, Jeschke, G., and Goldfarb, D. (2005) *J. Chem. Phys.*, **122**, 024515, 1-11. With permission, © 2005 American Institute of Physics.

3. In ESEEM experiments, the echo modulation depth tends to go to zero at high microwave frequencies/magnetic fields. In contrast, in ENDOR the nuclei are even better separated at high microwave frequencies/magnetic fields (based on their nuclear Larmor frequencies).

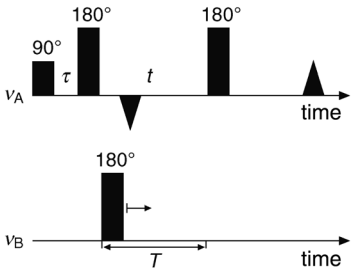


Figure 18.54 Four-pulse PELDOR sequence; ν_A and ν_B denote the two different microwave frequencies.

PELDOR – Pulsed Electron Double Resonance With the pulsed electron double resonance (PELDOR) sequence (Figure 18.54), the dipolar coupling between two unpaired electrons A and B can be selectively detected. Here, the detection sequence 90°-τ-180°-t-180° is applied to electron A using a microwave frequency ν_A. This leads to a refocused spin echo, which can be detected after the time interval t-τ after the last pulse. Within the time interval T, a pump pulse with the microwave frequency ν_B is applied, which inverts the spin of electron B in the same molecule. When the two unpaired electrons are coupled, the inversion of the electron B results in a change of the magnetic field at the electron A and thus a change in the echo amplitude. Moving the pump pulse within the time interval T leads to an oscillation of the observed echo amplitude.

The frequency of this oscillation is the frequency of the electron–electron coupling $\nu_{AB} = J + D$; J can usually be neglected for distances $r_{AB} > 1.5$ nm, which in turn makes ν_{AB} depend only on D. This dipolar frequency depends only on the distance between the two electrons as well as the orientation-term $(1 - 3\cos^2\theta)$ (Section 18.2.6). Since the measurement is performed in frozen solution and usually without orienting the sample, all orientations of θ are detected. This yields in the frequency domain the so-called dipolar Pake pattern from which the frequency for θ (90°) can be read off and with which the distance r_{AB} can be directly calculated.

With this pulse sequence, distances of up to 160 Å between two spin centers can be measured. An example is shown in Figure 18.55, where the time domain spectrum for a DNA molecule labeled with two nitroxides is presented. The figure also shows the frequency domain spectrum obtained by Fourier transformation (termed Pake spectrum or Pake pattern). The intense line at 6.8 MHz corresponds to the orientation $\theta = 90°$. From this value a distance r_{AB} of 19.5 Å can be

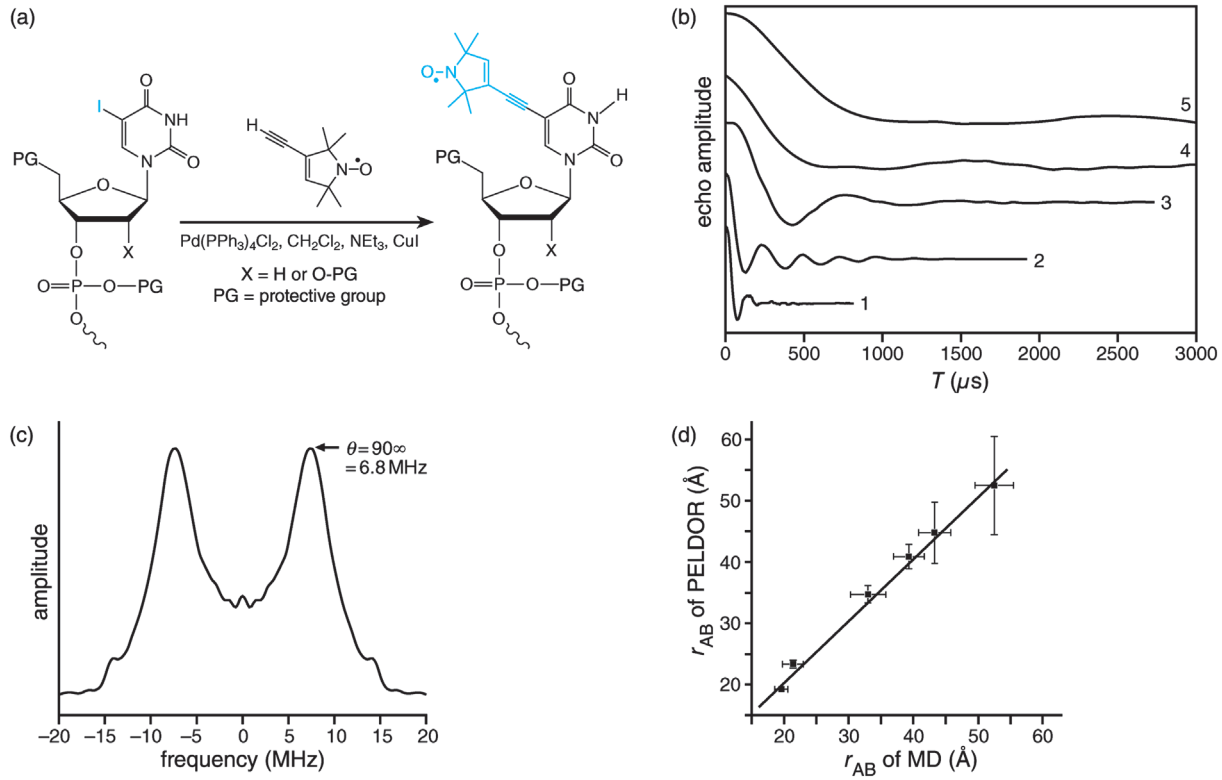


Figure 18.55 (a) Reaction scheme for the covalent attachment of nitroxides to DNA or RNA; (b) PELDOR – time domain spectra for a series of five DNA duplexes in which the distance between the two nitroxides increases from 1 to 5; (c) Fourier transform PELDOR spectrum (Pake spectrum) of DNA 1; (d) correlation of PELDOR distances and molecular dynamics simulations for a number of DNA and RNA duplexes. Source: Schiemann, O. *et al.* (2004) *J. Am. Chem. Soc.*, **126**, 5722–5731. With permission, Copyright © 2004, American Chemical Society.

calculated (Section 18.2.6), which matches very well with the theoretically expected distance of 19.3 Å.

Comparison between PELDOR and FRET

Another spectroscopic method, which can be used to measure distances in the nanometer range, is FRET (fluorescence resonance energy transfer). A comparison between FRET and PELDOR shows that the methods are complementary.

FRET, Section 16.7

- FRET provides distances of biomolecules in liquid solution. PELDOR, on the other hand, is performed in frozen solution.
- For FRET measurements a single molecule is sufficient, while concentrations in the micromolar range are required for PELDOR.
- FRET can observe distance changes in a time-resolved manner. PELDOR observes frozen distance distributions.
- In a PELDOR experiment, the coupling mechanism between both spin centers can be resolved and the size of J can be determined. In FRET experiments, the mechanism of fluorescence quenching is not always clear; often, reference measurements are needed.
- Calculation of the distance from the Pake spectrum is parameter-free. For the analysis of FRET measurements, assumptions about the orientation parameter κ must be made.
- Different labels are used for FRET and PELDOR. Large chromophores are frequently used in FRET. These are attached to the biomolecules via very flexible linkers. EPR labels are small and can be attached close to the surface, sometimes by rigid linkers (e.g., for DNA or RNA). Thus, it is easier to correlate the measured distances to the structure of the biomolecule. On the other hand, rigid linkers have a higher propensity to change the structure of the molecule under investigation. In reality, for both cases care should be taken to avoid the induction of structural changes.

18.2.8 Further Examples of EPR Applications

In the previous sections several examples were presented for the determination of structural elements by EPR methods. This section describes three examples for the determination of mobilities, pH values, and binding constants.

Quantification of Spin Sites and Binding Constants The intensity of the EPR signal depends, *inter alia,* on the number of electron spins in a sample. However, since many other factors influence the signal intensity, the spin number cannot be directly determined from the signal intensity in a straightforward manner. The number of spins can only be determined by comparison with a reference sample. For this, the unknown sample and the reference sample must be measured under exactly identical conditions. For technical reasons it is difficult to do so and thus the error for such experiments is normally about 15%. Nevertheless, in this way the number of spins per biomolecule can be determined if the concentration of the biomolecule is known. In addition, if the EPR spectra of a protein-bound and protein-free paramagnetic center are different, the number of binding sites, and the dissociation constants of the binding sites, can be determined. An example of this approach is described below, namely, the binding of Mn(II) to a catalytically active RNA, the minimal hammerhead ribozyme (Figure 18.56).

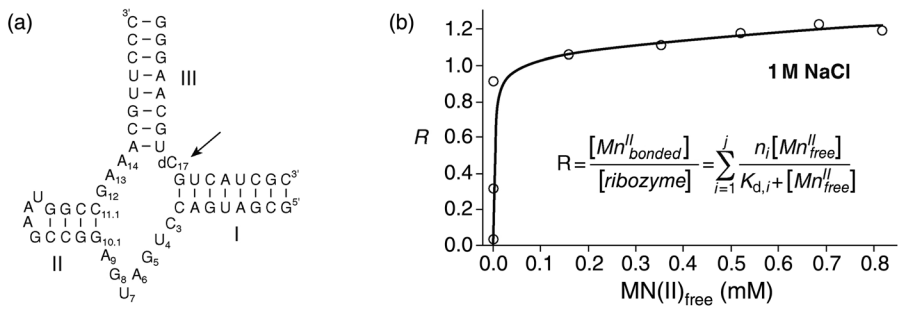


Figure 18.56 (a) Secondary structure of the minimal hammerhead ribozyme. The arrow indicates the cleavage site in the phosphodiester backbone. (b) Binding isotherm obtained from the EPR titration. The open circles are the experimental data while the solid line is the fit using the formula given in the graph. Source: Kisseleva, N. *et al.* (2005) *RNA*, **11**, 1–6. With permission, Copyright © 2005 by RNA Society.

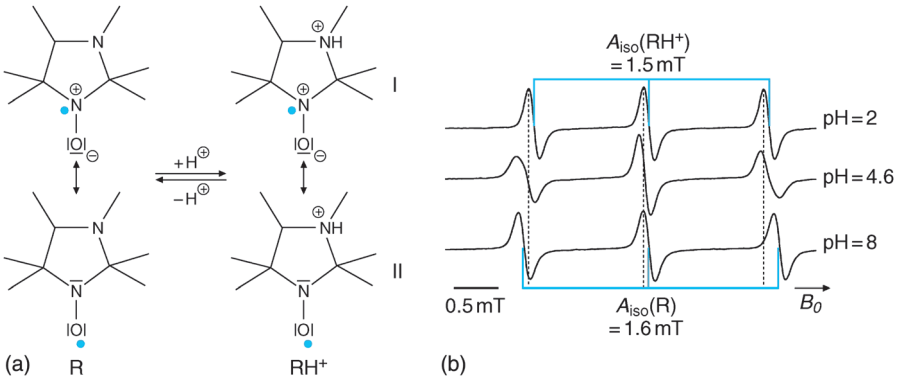


Figure 18.57 (a) Structure of the protonated/unprotonated nitroxide. In the protonated nitroxide, the mesomeric form I is energetically unfavorable due to the repulsion between the two positive charges. Thus, form II predominates, where the electron spin is located on the oxygen. (b) L-Band (1.3 GHz/40 mT) cw-EPR spectrum of the nitroxide at three different pH values. The difference in ^{14}N hyperfine coupling can be clearly seen. The pK_a value of the nitroxide is 4.6. Therefore, at pH = pK_a = 4.6 both forms are present in a 1 : 1 ratio. Source: Sotgiu, A. *et al.* (1998) *Phys. Med. Biol.*, **43**, 1921–1930. With permission, copyright © 1998 IOP.

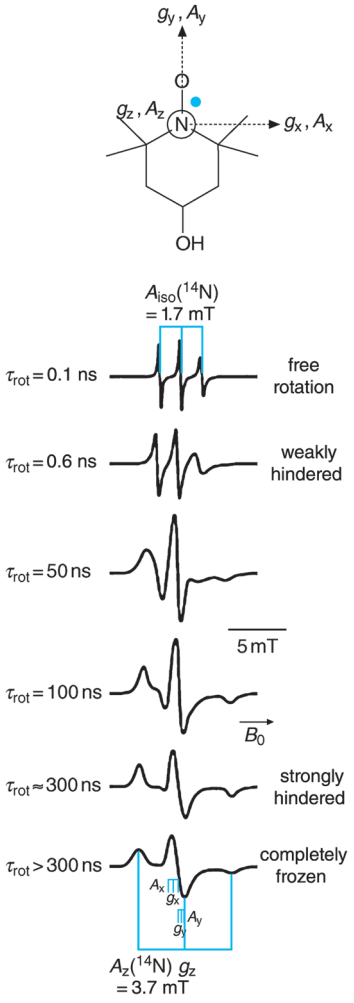


Figure 18.58 Influence of the rotational correlation time on the cw-X-band EPR spectrum of the nitroxide Tempol. In the case of free rotation, the three hyperfine lines of the nitroxide are split by A_{iso} = 1.7 mT. If the rotation is completely frozen, the nitroxide is dominated by the anisotropic hyperfine coupling constant A_z of 3.7 mT. Source: Weber, S., Wolff, T., and von Bünau, G. (1996) *J. Colloid Interface Sci.*, **184**, 163–169. With permission, Copyright © 1996 Academic Press.

When small quantities of Mn(II) are titrated to a buffer solution containing the minimum hammerhead ribozyme, a much smaller cw-EPR signal is obtained than if the same amount of Mn(II) is titrated into a buffer solution without ribozyme. The reason for this is that Mn(II) bound to the ribozyme yields very broad EPR lines, such that the EPR spectrum of the Mn(II)/ribozyme complex cannot be observed at room temperature. Thus, the concentration of bound Mn(II) can be calculated from the signal intensity originating from Mn(II) free in solution. By plotting the ratio of bound Mn(II) against the amount of free Mn(II), a binding isotherm is obtained. In this way, the dissociation constant for ribozyme-bound Mn(II) was determined to be 4 μM.

Local pH Values Nitroxides, which contain an amino group in or on the ring system, are often used as pH probes. The principle of such pH sensors is based on the protonation of the amino group in an acidic solution, whereby a positive charge is produced close to the nitroxide. The positive charge causes a shift in the spin density from the nitrogen to the oxygen atom, and thus a reduction of the ^{14}N hyperfine coupling and the *g*-tensor (Figure 18.57a). Depending on the pH, the cw-EPR spectrum is then a superposition of the spectra of the protonated and deprotonated nitroxide. From the intensity ratio of the two spectra, the concentration ratio of the two forms and thus the pH can be obtained. If such a nitroxide is bound to a biomolecule, the pH can be measured in the local environment of the biomolecule. A similar dependency of *A* and *g* on the polarity can be used to distinguish the membrane interior from the membrane surface.

Mobility Nitroxide spin labels are also frequently used to obtain information on the mobility of biomolecules. This method makes use of the fact that the spectrum of the nitroxide changes depending on its rotational freedom. If the nitroxide is free in its rotational movement, an isotropic three-line spectrum is obtained at X-band (Figure 18.58, upper spectrum). If the rotation is completely frozen, one observes an anisotropic spectrum as shown in Figure 18.58 (bottom) (also compare with nitroxide at 180 GHz, Figure 18.42). Depending on the degree of rotational freedom of the nitroxide, the spectrum changes gradually from isotropic to anisotropic.

A measure of the rotational freedom is the rotational correlation time τ_{rot}, which can be easily calculated from the EPR line shapes and intensities:

$$\tau_{rot} = 6.5 \times 10^{-10} \Delta B \left(\sqrt{\frac{h_0}{h_1}} - 1 \right) \qquad (18.36)$$

Here, h_0 is the intensity of the central line, h_1 is the intensity of the low-field line, and ΔB is the width of the central line in Tesla. The Tesla is the unit for the external magnetic field $\boldsymbol{B}$ (actually the magnetic induction); 1 T = 10^4 Gauss.

If a nitroxide is covalently bound to a biomolecule, the freedom of rotation of the former is limited by the freedom of rotation of the latter. The measured rotational correlation time of the nitroxide is thus a measure of the mobility of the biomolecule. However, it is difficult to separate the τ_{rot} value of the biomolecule from the measured τ_{rot} value, which has a residual contribution from the mobility of the label independent of the host biomolecule. It is simpler to determine

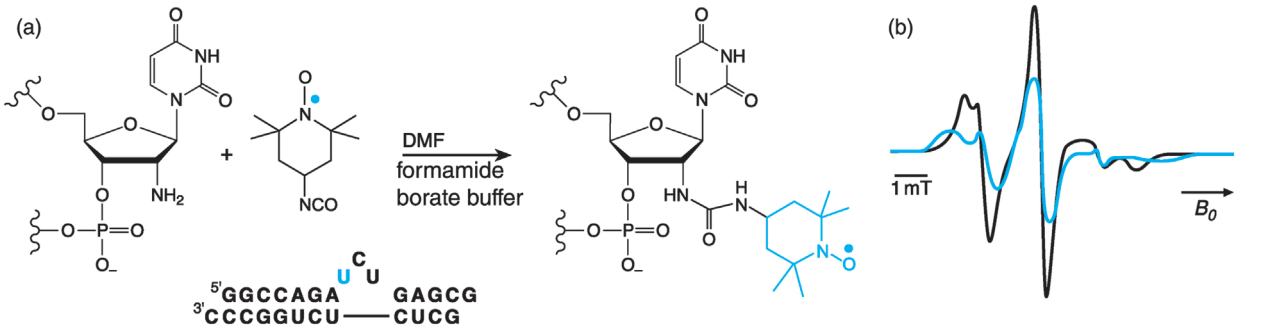


Figure 18.59 Binding of the TAT protein to the nitroxide-labeled HIV TAR RNA. (a) Reaction scheme for the labeling of RNA with a nitroxide (top) and the secondary structure of the TAR RNA with the nitroxide located at the blue uridine (below). (b) The cw-X-band EPR spectra of TAR RNA spin-labeled at U23 without (black) and with (blue) the bound TAT protein. The broadening of the spectrum and the decrease in the intensity of the low-field line for the TAR-TAT-complex can be clearly seen. By the analysis of multiple spectra, in which the nitroxide is bound to different positions on the RNA, statements about the influence of the dynamics of the RNA on the binding of the protein could be made. Source: Edwards, *et al*. (2002) *Chem. Biol.*, **9**, 699–706. With permission, Copyright © 2002 Cell Press. Published by Elsevier Ltd.

relative differences. The binding of ligands to RNA can be followed in this way by EPR spectroscopy (Figure 18.59).

18.2.9 General Remarks on the Significance of EPR Spectra

In most cases it is difficult to derive a clear statement or structure from a single EPR spectrum. To obtain reliable results, one can proceed as follows:

- vary the sample conditions (e.g., temperature, solvent, etc.);
- modify the biomolecule biochemically (e.g., protein mutants, spin label positions, isotope labeling);
- record cw-EPR spectra at different microwave frequencies to resolve g-values and spectra of various radicals; this also allows the hyperfine coupling contribution to be separated from g-tensor contributions;
- combine several pulse-EPR/ENDOR-methods to select and assign individual spectral hyperfine contributions;
- simulate the EPR spectra, to obtain the EPR parameters;
- translate the EPR parameters into structural data by comparison with results from quantum chemical methods (such as DFT), model systems, and literature data;
- combine with further spectroscopic methods.

18.2.10 Comparison EPR/NMR

As a final point, the two complementary magnetic resonance methods EPR and NMR are compared.

- NMR spectroscopy investigates the spin of magnetic nuclei in predominantly diamagnetic samples and because protons, nitrogen, or carbon nuclei occur everywhere in the biomolecule the method can elucidate the overall structure of the investigated biomolecule on an atomic level. EPR spectroscopy on the other hand detects the spin of unpaired electrons (paramagnetic samples) and observes only the local environment of the spin center.
- This local restriction of EPR also means that there is no size restriction for the biomolecule to be studied, while NMR is currently limited to biomolecules with a mass of roughly 80 KDa.
- The sensitivity of EPR spectroscopy is higher (nanomol) than that of NMR spectroscopy (millimol). This is due to the larger magnetic moment of the electron ($\mu_e/\mu_H = -1838$), which makes the Boltzmann population difference larger for EPR (1.1×10^{-3} compared to 1.1×10^{-6} at 3.4 T and $T = 300$ K).
- Due to the larger magnetic moment, the relaxation processes are faster in EPR. This is why the time scale for pulsed EPR experiments is nano- to microseconds as opposed to

milliseconds to seconds for NMR. This and the need for the use of microwaves also mean that the technical requirements are higher for EPR.
- The faster relaxation leads to broader lines in EPR (MHz versus Hz).
- The electron–nucleus spin coupling in EPR is larger than nuclear–nuclear spin coupling in NMR due to the greater magnetic moment of the electrons (MHz versus Hz). For this reason, larger distances between the unpaired electron and a nucleus (to 10 Å) or the electron and another unpaired electron (up to 80 Å) can be determined by EPR.
- The theoretical effort needed for the simulation of EPR spectra is significantly larger than for high-resolution liquid state NMR spectra due to the anisotropic contributions and the very fast relaxation.
- The translation of EPR and also NMR parameters into structural conclusions using quantum chemical calculation methods (such as DFT) is still complicated. This is particularly significant for transition metals for which often only trends are obtained.
- Both NMR and EPR can be carried out in either liquid buffer solutions or in frozen solutions/powders. This is to say that no single crystals are needed for either method.
- Both spectroscopic methods can provide insight into the dynamics of biomolecules.

Acknowledgements

We thank Yaser NejatyJahromy for careful reading of the manuscript.

Further Reading

Section 18.1

Bax, A. (2003) Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Sci.*, **12**, 1–16.
Blumenthal, L.M. (1970) *Theory and Application of Distance Geometry*, Chelsea, Bronx, New York.
Cavanagh, J., Fairbrother, W.J., Palmer, A.G. III and Skelton, N.J. (1996) *Protein NMR Spectroscopy*, Academic Press.
Creighton, T.E. (ed.) (1992) *Protein Folding*, Freeman, New York.
Croasmun, W.R. and Carlson, R.M. (eds) (1994) *Two-Dimensional NMR Spectroscopy*, VCH-Verlagsgesellschaft, Weinheim.
Derome, A.E. (1987) *Modern NMR Techniques for Chemistry Research*, Pergamon, Oxford.
Dingley, A.J., Cordier, F., and Grzesiek, S. (2001) An introduction to hydrogen bond scalar couplings. *Concepts Magn. Reson.*, **13**, 103–127.
Ernst, R.R. (1992) Kernresonanz-Fourier-transformationsspektroskopie (Nobel-Vortrag). *Angew. Chem.*, **104**, 817–952.
Ernst, R.R., Bodenhausen, G., and Wokaun, A. (1987) *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*, Clarendon Press, Oxford.
Evans, J.N.S. (1995) *Biomolecular NMR Spectroscopy*, Oxford University Press.
Fernandez, C. and Wider, G. (2003) TROSY in NMR studies of the structure and function of large biological macromolecules. *Curr. Opin. Struct. Biol.*, **13**, 570–580.
Friebolin, H. (1998) *Ein- und Zweidimensionale NMR-Spektroskopie*, VCH-Verlagsgesellschaft, Weinheim.
Goldman, M. (1988) *Quantum Description of High-Resolution NMR in Liquids*, Clarendon Press.
Karplus, M. and Petsko, G.A. (1990) Molecular dynamics simulations in biology. *Nature*, **347**, 631–639.
Pain, R.H. (ed.) (1994) *Mechanisms of Protein Folding*, Oxford University Press, Oxford.
van de Ven, F.J.M. (1995) *Multidimensional NMR in Liquids*, VCH-Verlagsgesellschaft, Weinheim.
Van Gunsteren, W.F. and Berendsen, H.J.C. (1990) Moleküldynamik-computersimulationen: methodik, anwendungen und perspektiven in der chemie. *Angew. Chem.*, **102**, 1020.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley & Sons, Inc., New York.
Weltner, W. (1983) *Magnetic Atoms and Molecules*, Dover Publications, New York.

Section 18.2

Atherton, N.M. (1993) *Principles of Electron Spin Resonance*, Ellis Horwood, New York.
Berliner, L.J. (ed.) (1998) *Spin Labeling: The Next Millennium, Biological Magnetic Resonance*, vol. 14, Kluwer Publishing, Amsterdam.

Dikanov, S.A. and Tsvetkov, Y.D. (1992) *Electron Spin Echo Envelope Modulation (ESEEM) Spectroscopy*, CRC Press, Boca Raton, FL.
Eaton, G.R., Eaton, S.S., and Berliner, L.J. (eds) (2000) *Distance Measurements, Biological Magnetic Resonance*, vol. 19, Kluwer Publishing, Amsterdam.
Kaupp, M., Bühl, M., and Malkin, V.G. (eds) (2004) *Calculation of NMR and EPR Parameters*, Wiley-VCH Verlag GmbH, Weinheim.
Poole, C.P. (1983) *Electron Spin Resonance – A Comprehensive Treatise on Experimental Techniques*. Wiley-Interscience, New York.
Schweiger, A. and Jeschke, G. (2001) *Principles of Pulse Electron Paramagnetic Resonance*, Oxford University Press, Oxford.
Misra, S.K. (ed.) (2011) *Multifrequency Electron Paramagnetic Resonance*, Wiley-VCH Verlag GmbH, Weinheim.
Weil, J.A., Bolton, J.R., and Wertz, J.E. (1994) *Electron Paramagnetic Resonance: Elementary Theory and Practical Applications*, Wiley-Interscience, New York.

Electron Microscopy

19

Harald Engelhardt

Max-Planck-Institut für Biochemie, Am Klopferspitz 18, 82152 Martinsried, Germany

Modern microscopic techniques produce images of small organisms, tissues, single cells, organelles, membranes, macromolecular assemblies, isolated macromolecules, and of small molecules down to atoms (Figure 19.1). The beginning of microscopy dates back to the seventeenth century when Antoni van Leeuwenhoek (1632–1723) in the Netherlands and Robert Hooke (1635–1703) in England built their first simple instruments and initiated the development of light microscopy.

Light Microscopy, Chapters 7, 8

The scientific investigation of optical systems and of its resolution limit by Ernst Abbe (1840–1905) in Germany, the improvement of the microscopic illumination by August Köhler (1866–1948), the development of better glass materials, and Robert Koch's famous microbiological investigations led to a blooming of microscopy in science. In the 1930s, the Dutch physicist Frits Zernike (1888–1966) devised the phase contrast microscope, which allowed

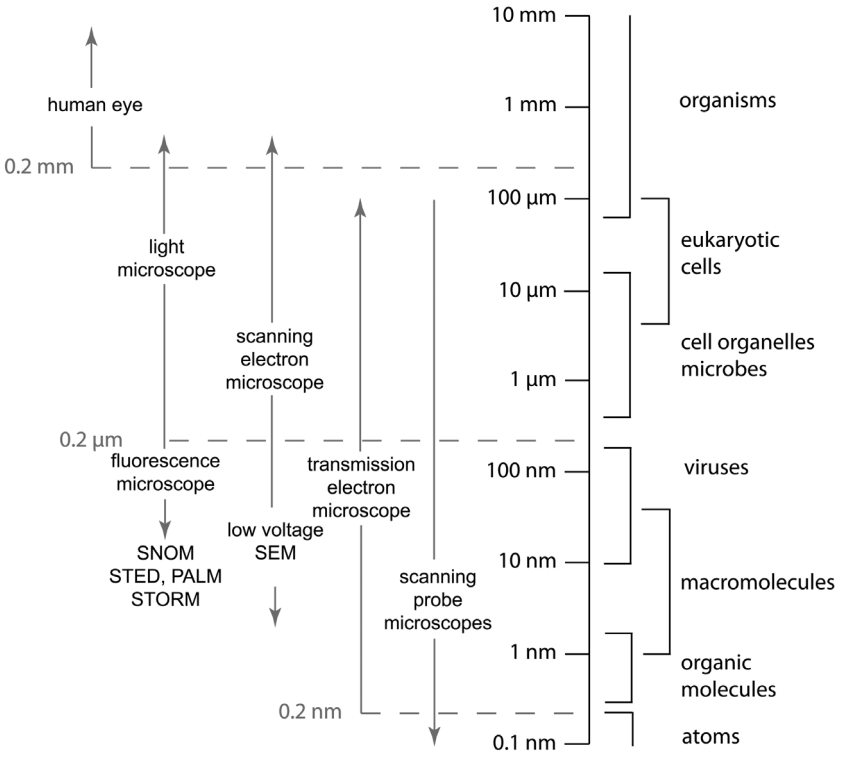


Figure 19.1 Biological structures and suitable microscopy techniques. The resolution limits of the human eye, light microscopes, and the transmission electron microscope are indicated. Fluorescence microscopes show the position of fluorophores where the optical near field microscopies (e.g., SNOM), the stimulated emission depletion (STED), photoactivated localization microscopy (PALM), and the stochastic optical reconstruction microscopy (STORM) overcome the classical resolution limit of light microscopes. Scanning electron and scanning probe microscopies provide surface structure information, light and transmission electron microscopy allow three-dimensional imaging.

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

researchers to investigate unstained cells and tissue of low contrast and whose basic principle also applies to electron microscopy.

Köhler also introduced the first fluorescence microscope (1908), an instrument that gained more significance in cytological research after the development of stronger light sources. Based on the increasing success of molecular genetics and the fusion of the green fluorescent protein to other macromolecules it became possible to identify and localize cellular protein complexes, and this success enormously stimulated the field of cellular biology in the second half of the 1990s. At this time, fluorescence microscopy experienced another revolutionary breakthrough by overcoming the classical resolution limit that the refraction law and the wavelength of light dictate. The new techniques allow the detection of single fluorescence markers in cells with a spatial accuracy of about 30 nm or better. Several instrumental solutions were introduced, such as STED (stimulated emission depletion microscopy), invented by Stefan Hell (*1962), or PALM (photoactivated localization microscopy), and STORM (stochastic optical reconstruction microscopy). The latter stimulate fluorescence labels locally and calculate the center of the fluorescence signal afterwards. None of these fluorescence microscopes provides genuine structural information, yet, they locate labeled molecules and display their distribution with unprecedented spatial precision (super-resolution).

Atomic Force Microscopy, Chapter 20

The scanning probe microscopies (SPM) in the optical near-field (scanning near-field optical microscopy, SNOM; scanning near-field infrared microscopy, SNIM) are also techniques that detect objects of "submicroscopic" dimensions and can even record spectroscopic information in the nanometer range independently of the wavelength of the light. The field of "nano-optics" has important applications in material science. The "apertureless" microscopes make use of scanning force microscopy (SFM; atomic force microscopy, AFM) that was introduced by Gerd Binnig (*1947) in 1986, only a few years after he and Heinrich Rohrer (1933–2013) had invented the scanning tunneling microscope (STM), a novel type of microscopy for imaging surfaces. Several variants of SFM are also useful for biological structure research and are complementary to applications in light and electron microscopy.

When Louis de Broglie (1892–1987) realized that electrons can be understood as waves with wavelengths far below one nanometer, he determined the theoretical basis for a microscope operating with electrons. Max Knoll (1897–1969), Ernst Ruska (1906–1988), and Bodo von Borries (1905–1956) developed the first transmission electron microscope (TEM) in 1931, and only two years later they obtained images with much higher resolution than light microscopy. Today it is possible to resolve single atoms in radiation-resistant objects such as alloys. Manfred von Ardenne (1907–1997) built the first scanning electron microscope (SEM) in 1937, an instrument that produces impressive surface images of high depth of field.

Biological objects such as proteins that are very radiation-sensitive require preparative and technical efforts before imaging with quasi-atomic resolution ($\leq$0.3 nm). The two-dimensional (2D) crystalline bacteriorhodopsin from halobacteria and the photosynthetic light-harvesting antenna complex LH_{II} of the photosynthetic membranes from chloroplasts were the first biological specimens whose three-dimensional (3D) atomic structures were solved by means of electron microscopy (electron crystallography).

X-Ray Strucrure Analysis, Chapter 21

In addition to X-ray crystallography and NMR spectroscopy, electron microscopy is the third method with which we can determine the spatial structure of macromolecules. Electrons interact particularly strongly with the object and, in contrast to the other methods, also provide images of single molecules. Automated data acquisition and novel electron detectors enable us now to collect data of single protein complexes of sufficient quantity and quality to achieve quasi-atomic resolution. Single particle electron microscopy is indeed the only way to resolve the 3D structure of macromolecular complexes that are too big for NMR spectroscopy and too flexible for X-ray crystallography and do not form crystals.

NMR Spectroscopy, Chapter 18

Wolfgang Baumeister (*1946) and his team developed an approach that made the 3D structure of native, shock-frozen (vitrified) cells accessible to TEM at macromolecular resolution. The 3D reconstruction of the cytoskeleton in intact cells from the slime mold *Dictyostelium discoideum* signaled the breakthrough of cryo-electron tomography (CET) in 2002. Meanwhile, CET has granted unprecedented insights into the macromolecular organization of microorganisms and eukaryotic cells and has opened up new and exciting perspectives in structure research *in situ*. In 2014 Baumeister's laboratory introduced a new tool (phase plate) for phase contrast enhancement in electron microscopy that has the potential to become standard equipment. Cryo-electron tomography now integrates

molecular and cellular structure research, which before were separate biological disciplines.

19.1 Transmission Electron Microscopy – Instrumentation

TEM received its name from the imaging mode. The electrons that transmit the object are used for imaging, similar to the function of light in optical microscopy. The basic construction of electron and optical microscopes are indeed similar, except that electron magnetic lenses are used instead of glass lenses and the electron microscopes are much bigger (Figure 19.2). The lenses consist of iron-sheathed coils that create a magnetic field upon current flow and direct the field towards the inner space. Electron magnetic lenses are always converging lenses. Correcting spherical and chromatic aberrations is thus challenging, yet, in recent years, correction systems have improved the image formation in high-resolution electron microscopes in particular. The electron source is a cathode that emits electrons. An anode with a potential difference to the cathode of 10^5 V or more accelerates these electrons. The condenser lens (often two lenses after each other) focuses the electron beam in the object plane where it transmits the object and is expanded by the object lens. Its focus length can be adjusted by variation of the lens current so that changing the magnification does not require other objective lenses. One or more projection lenses between the object and the imaging plane further expand the beam and lead to an increase in magnification. The image is rendered visible on a fluorescence screen (which may be inspected by a binocular loupe with tenfold magnification) and recorded on film plates in the conventional way or, nowadays, by means of special digital cameras. The microscopes allow for magnification between ≤100-fold and about 10^6-fold. However, for macromolecules and biological structures a primary magnification of ≤100 000-fold is sufficient. The electron microscope column must be evacuated so that electrons only interact with object atoms but not with gas molecules. Microscopy in a vacuum necessitates preparation of the biological objects (fixation, dehydration, embedding in plastic material or freezing for

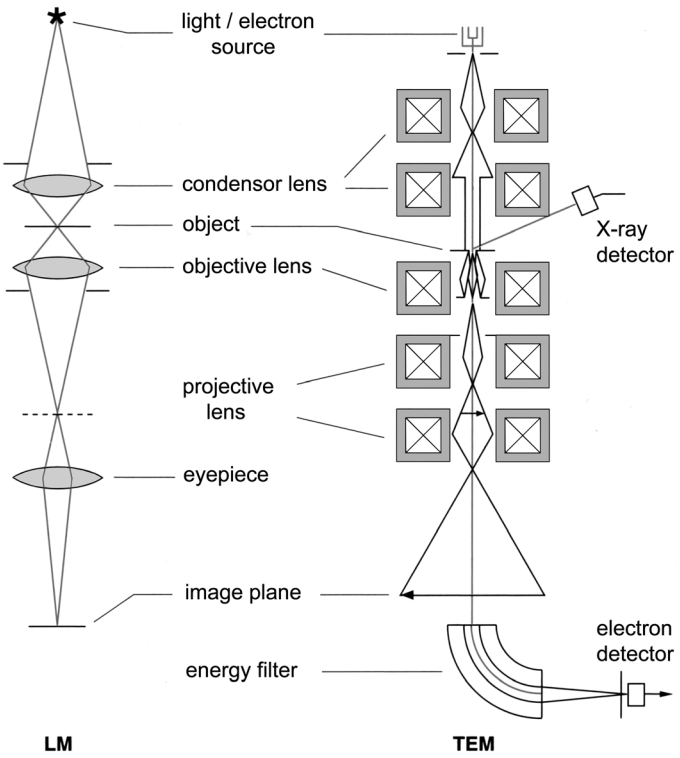


Figure 19.2 Beam path in the light microscope (LM) and in the transmission electron microscope (TEM). The principal system of lenses is the same in both microscopes, the TEM usually contains a second condenser and projection lenses in addition. Condenser and objective apertures limit the illumination and blank out strongly diffracted regions of the beam, thereby enhancing the contrast. The image is observed in the image plane by eye (LM) or on a fluorescence screen (TEM) and can be recorded by a camera. An external energy filter separates electrons of different energy and enables electron energy loss spectroscopy (EELS) analysis and electron spectroscopic imaging (ESI) (see also Figures 19.9 and 19.12). X-Rays that are created by interaction of beam electrons with object electrons are either shielded or can be recorded by a detector.

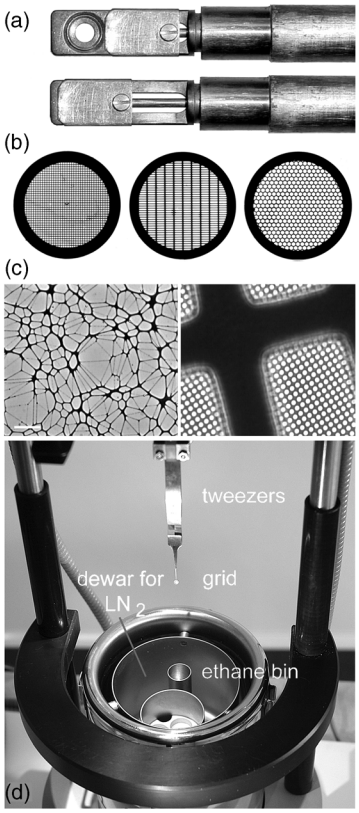


Figure 19.3 Object holder, grids, and plunger for biological cryosamples. (a) Object holder with a mechanical shield to protect the frozen sample from contaminations. The holder can be rotated around its longitudinal axis by about ±70° to record different projections of an individual object for 3D reconstruction (tilt series). (b) Grids (diameter 3 mm) of various designs for TEM. They are usually covered with a 5–10 nm thick carbon film for small biological specimens. (c) Carbon films with holes for cryosamples. A thin ice film containing the biological structures spans the holes. Bar indicates 20 µm. (d) Plunger for vitrification of biological samples for cryo-electron microscopy. Tweezers hold the grid and inject it into a liquid cryogen (ethane) cooled by liquid nitrogen to about −180 °C. The high cooling rate vitrifies water and prevents the growth of ice crystals. The biological structures are preserved in a close-to-live state.

cryo-electron microscopy). If not doing so, the specimens would dehydrate in the microscope and become damaged, and the vacuum would break down.

The specimens are placed on copper grids (sometimes made of gold or nickel) with meshes of 30–100 µm (Figure 19.3). The grids are covered by a thin (5–10 nm) carbon film, which supports isolated macromolecules or ultrathin sections. These films are deposited on surfaces of freshly broken mica by heat evaporation of graphite in a vacuum chamber. The film is then floated onto the surface of water and transferred to the grids. For cryo-electron microscopy these films contain holes of 0.5–2 µm in diameter. Comparable grids are commercially available ready for use (Lacey® and Quantifoil® grids; Figure 19.3b). Very recently, we saw the introduction of gold grids with regularly perforated (1 µm) gold membranes instead of carbon films; these are extremely stable in the electron beam and work well for high-resolution imaging. The holes are filled by water, which becomes thin ice films after freezing, and they thus contain the biological objects (macromolecules, cellular components, or organelles, cells). Carbon-coated grids are hydrophobic and have to be made hydrophilic prior to the application of material dissolved or suspended in water. Ionized gas molecules – produced in evacuated chambers with glow discharge (plasma cleaner) – render carbon film temporarily wettable. As a side effect, the plasma destroys impurities and cleans the grid surface.

The grids are mounted on an object holder and inserted into the microscope (Figure 19.3a). Frozen specimens that are examined in the cold (−180 °C or less, cryo-electron microscopy, see Sections 19.2.1 and 19.3.3) must be continuously cooled by liquid nitrogen in a controlled manner to avoid recrystallization of the amorphous ice above −150 °C. The object holder can be tilted around one axis or different axes (ideally being set perpendicularly to the first one), and the specimen can be inspected from different projection angles, which is necessary for 3D-reconstruction in electron tomography (Section 19.5).

19.2 Approaches to Preparation

Biological specimens must be sufficiently thin so that the strongly interacting electrons can pass through them. Intact cells or tissue preparations are usually chemically fixed, dehydrated, and embedded in special resins (*Epon*®) or in material that polymerizes in the cold (e.g., *Lowicryl*®). Ultrathin sections (≤100 nm) are cut from these polymer blocks by means of ultramicrotomes. Fixation, staining, and thin sectioning has been the standard procedure in cell biology for about five decades, and we owe most of our knowledge of cell architecture to this technique. We will not go into further details here since shock-frozen and untreated biological samples are currently replacing chemically fixed and stained specimens. The cryotechniques preserve the native structure and organization of macromolecules in the cellular context and open up new perspectives in cytological research (Figure 19.4). Section 19.2.1 describes the preparation of intact cells in amorphous ice and various thinning approaches for cryo-electron microscopy. Electron microscopy of isolated cell wall fragments, membranes, proteins, and other macromolecular complexes does not require thin sectioning; these objects are already thin enough. In addition to cryopreparation, Sections 19.2.2–19.2.4 describe common contrasting and labeling procedures that are used for a quick inspection of soluble macromolecules or for special applications.

19.2.1 Native Samples in Ice

If the inner structure of molecules, of macromolecular assemblies, or of intact cells is to be investigated, the object itself must be imaged and not the distribution of staining material. We therefore avoid any chemical fixation and contrast enhancement and look at the native object in aqueous solution that has been physically "fixed" by rapid freezing. Thin objects are directly applied to the grid, blotted to remove excess water until only a thin film is left, and shock-frozen by plunging the grid into a cryogen (liquid ethane or ethane–propane; Figure 19.3d). The high cooling rate ($\approx 10^5$ K s^{-1}) prevents water from forming ice crystals that would destroy the biological object. The ice remains amorphous; it is vitrified in a similar way as cold molten glass, as Jacques Dubochet (*1941) showed in 1981. To control conditions such as temperature

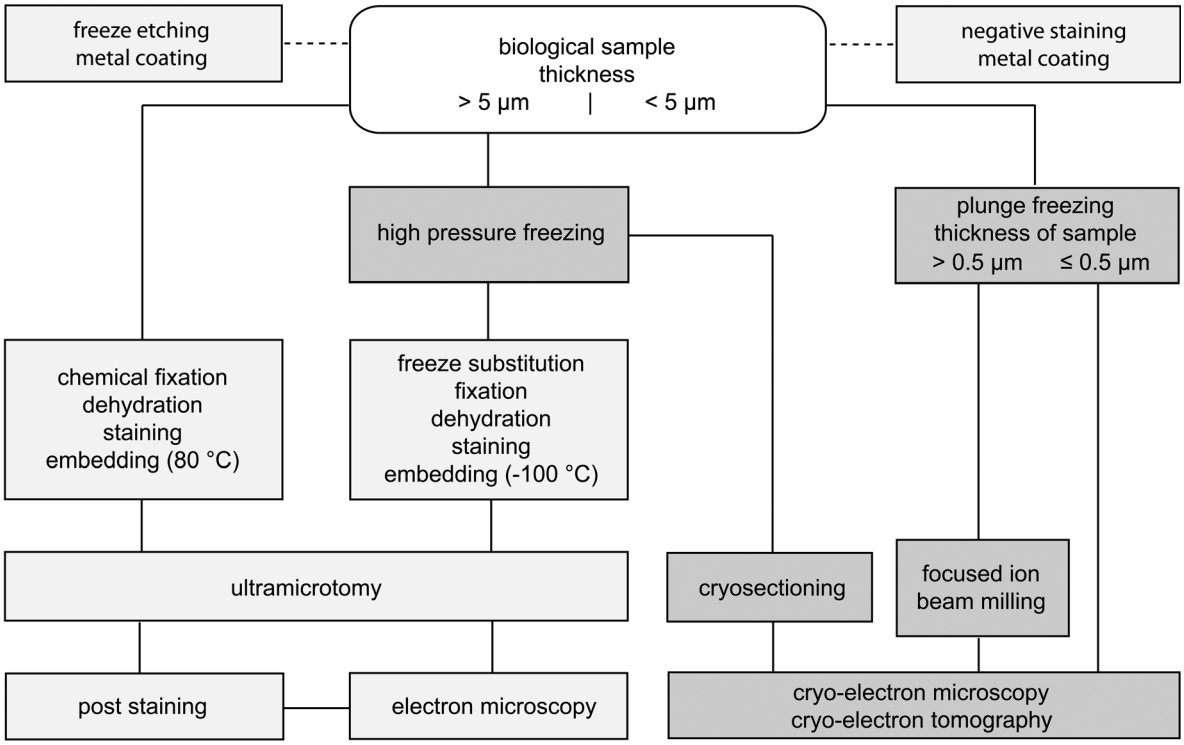


Figure 19.4 Preparation of biological samples. Cells and tissues are either treated by chemical fixation, embedding, and ultrathin sectioning for electron microscopy or for enhanced structural integrity by high-pressure freezing and freeze substitution. A close-to-live preparation is accomplished by vitrification, cryosectioning, or focused ion beam (FIB) milling. Thinner samples (cells, organelles, or isolated macromolecular complexes) are either thinned in the FIB or directly imaged in the TEM. Procedures for vitrified specimens are shaded in gray. The methods of freeze fracturing and negative staining are indicated but are not complete.

or light intensity, one can use a plunger with an integrated incubation chamber for automated blotting. Isolated protein complexes may be re-suspended in solutions of glucose or similar compounds (trehalose, tannic acid) if they have to be stabilized. The frozen grid is transferred to the microscope sample holder under liquid nitrogen and is imaged at −180 °C or less (Section 19.3.3). Since the contrast of protein (specific density ≈1.4 g cm^{-3}) in ice (≈1 g cm^{-3}) is low, it is advantageous to use grids with carbon films containing holes in order not to obliterate the weak contrast with other material (Figure 19.3c). However, cryopreparation of isolated protein complexes is standard and has replaced negative staining specimens for 3D structure determination.

Some eukaryotic cells, bacteria, archaea, and viruses as well as macromolecular complexes can be vitrified on grids and in most cases also directly inspected in a microscope (Figure 19.4). However, eukaryotic and multicellular specimens are often too thick to be vitrified by simple plunging. The heat transfer in the center of bigger samples (thickness >10 µm) is too slow to prevent the crystallization of water. However, solutions under pressure crystallize less quickly and the slower cooling rate suffices to vitrify samples 200–300 µm thick. High-pressure freezing instruments generate up to 0.2 GPa (≈2000 atm) at liquid nitrogen temperature. Moreover, the cell suspensions can be supplemented by anti-freeze compounds such as dextran, which is osmotically inert and reduces the ability of water to form ice crystals. The drawback, however, is that these additives mask polysaccharides of the cell surface.

Cryofixed cells or tissue may be freeze-fractured, freeze-etched, and metal-coated (replica technique; Figure 19.4) to investigate the cell surface or fracture faces. Alternatively, they can be fixed at about −100 °C, dehydrated, stained, and embedded in material that polymerizes in the cold. In cases were chemical fixation should be avoided, vitrified cells can be left untreated and are sectioned in the cryo-ultramicrotome.

The cell material is frozen in a small copper tube and this is mounted on a cryo-ultramicrotome in an atmosphere of liquid nitrogen (≈ −150 °C). The tube is trimmed so that the ice block is free

Vitrification Glass (*vitrum*) is an amorphous material that does not form crystals when it solidifies after melting. Water molecules on the other hand form ice crystals whose particular structure and density depend on the temperature and pressure during the freezing process. Currently, we know of 19 different ice forms. Very fast freezing to temperatures below -140 °C at normal pressure lets water assume an amorphous, vitrified state with a density of 0.94 g cm^{-3} (low density amorphous ice, **LDA**). The other types of amorphous ice have a rather high density (1.17 and 1.26 g cm^{-3}) and cannot be generated from liquid water. LDA is more similar to liquid water than any other form of ice.

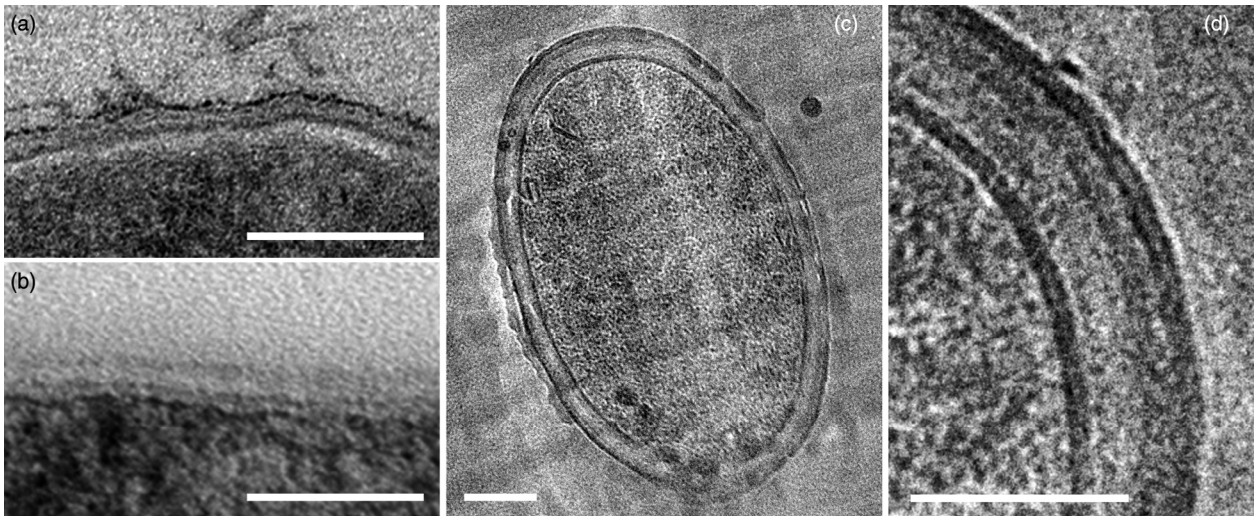


Figure 19.5 Ultrathin sections of *Mycobacterium smegmatis*. Preparation by (a) conventional fixation, dehydration, and embedding in epoxide resin (Epon©), (b) high-pressure freezing, freeze substitution, and embedding in Lowicryl©, and (c) high-pressure freezing and cryomicrotomy without any chemical treatment, (d) a higher magnification of (c). The lipid bilayer is only visible in cryosections; these are compressed in the cutting direction so that the originally round cross section of the bacterial cell becomes oval. Scale bars indicate 100 nm (a)–(c) and 50 nm (d). Source: parts (a) and (b) courtesy of Christopher Bleck, Basel, Switzerland.

from the surrounding copper and thin sections can be produced. Cryosectioning is not a routine approach, and it consists of up to 30% compression of samples in the sectioning direction. Despite several artifacts, cryosections provide insight into details of the cellular architecture that are usually lost during dehydration and plastic embedding (Figure 19.5).

A new development for thinning of frozen biological material is focused ion beam (FIB) micromachining. The vitrified samples are mounted in a scanning electron microscope (SEM) that is also equipped with an ion gun (gallium). The focused ion beam removes material from the surface of the object until it is thin enough for imaging in the TEM (300–500 nm). To save time and gallium, for ion milling the specimens should not be thicker than 5–10 µm. Appropriately thinned specimens are free of artifacts and will not be deformed.

An alternative approach ablates thin layers of biological material and images the surface of the sample block by SEM. When this procedure is repeated multiple times, the series of images, aligned and consecutively put together produce a 3D cube of the object. This technique is called FIB-SEM or "slice and view" and can be applied to cells and cell assemblies. Another approach uses an integrated microtome to smooth the surface instead of an ion beam; this is particularly applied to large embedded specimens such as tissues (mouse brain).

19.2.2 Negative Staining

Negative staining by heavy metal salts is a very simple and quick method to image isolated proteins, fibrillar assemblies, membranes, and similar objects at a resolution of about 2 nm. Negative staining is readily used to inspect preparations in terms of purity and homogeneity and to get a first impression of the object structure.

A droplet of the sample (2–5 µl) is applied to a carbon-coated grid that has been made hydrophilic by glow discharge. After 15–60 s most of the liquid is blotted and the grid is washed with pure water, buffer, or salt solution ($\leq$10 mM) and stained by means of a heavy metal salt. Common stainings are $\approx$2% (w/v) solutions of uranyl acetate, phosphotungstic acid, or ammonium molybdate, amongst others. The compounds differ in contrast, radiation sensitivity, the applicable pH range, and their ionic characteristics. The metal salt covers the surface, fills holes and indentations of the macromolecules, and it is the distribution of the metal that is finally imaged (Figure 19.6). The metal coat is much more radiation-resistant than the biological material and preserves the spatial structure after drying with moderate irradiation. Negatively stained specimens may be stored for weeks to months. However, negative staining is usually not suitable for whole cells and bigger objects.

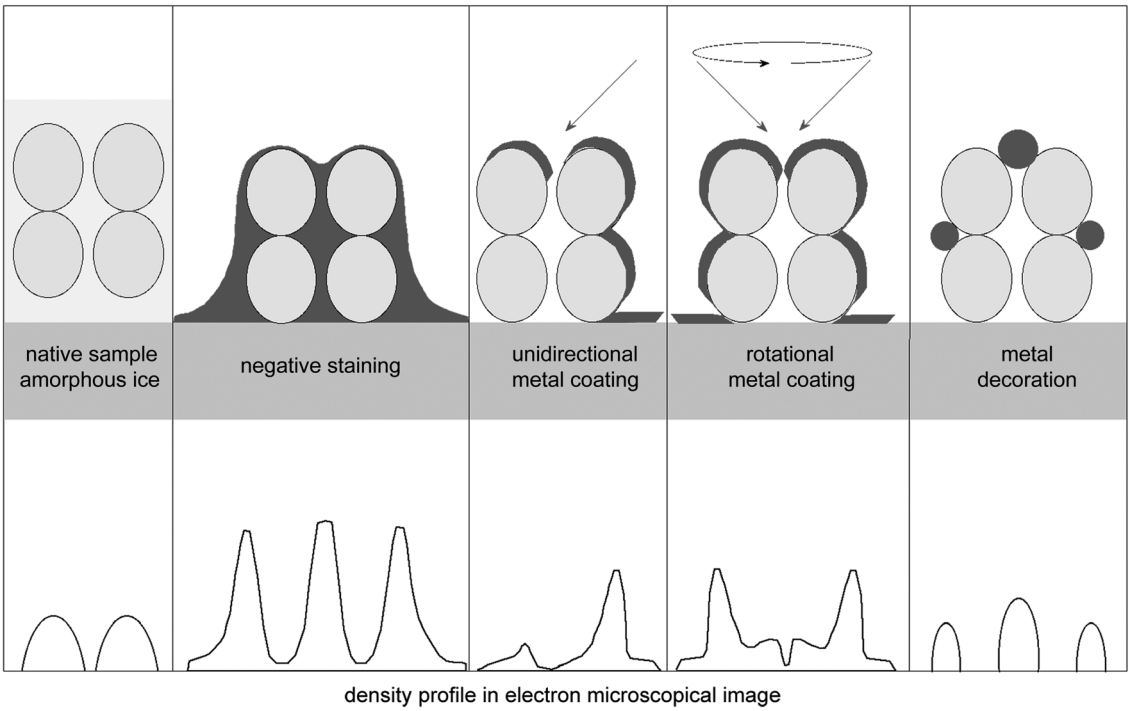


Figure 19.6 Schematic illustration of different contrasting methods for imaging macromolecules in the transmission electron microscope. Contrasting with a heavy metal leads to different density distributions in the EM image and provides structural information on individual components of a sample. Only images of ice-embedded native specimens show contrast originating from the object itself.

19.2.3 Metal Coating by Evaporation

A common method used to investigate intact cells or tissues in a microscope is by freezing, cutting, or fracturing them in the vacuum, sublimation of ice by freeze etching (at −80 °C), and contrasting the surface by evaporating heavy metal at an angle of 30–60°. The metal coat (1–2 nm Pt/C or other) is stabilized by 10–20 nm carbon (90°). Treatment with aggressive acid solutions removes the biological material, and only the remaining replica is inspected in the microscope. The grain size of the evaporated metal limits the resolution to about 2 nm, but complexes of macromolecules in membranes and cellular surfaces are detectable. Freeze fracturing was a common method in cytology and was used to obtain spatial information on cellular surfaces. The approach lost its importance with the introduction of cryo-electron tomography in structural research (Section 19.5.3).

Direct metal evaporation onto membranes or isolated protein assemblies that are adsorbed on carbon-coated grids is also possible. Since air-drying would destroy the unprotected biological specimens, they are freeze-dried to gently remove water and then contrasted in the cold, as in the freeze-etching approach. The metal is only deposited on the surface, pointing towards the evaporation source, and it renders the other regions invisible (Figure 19.6). The orientation of membranes or regular objects (2D crystals or S-layers) can thus be evaluated. Images of unidirectional metal-coated objects must be interpreted in a different manner than negatively stained samples since the density distribution resembles a landscape with hills and valleys that is transformed into a pattern of light and shadow ("metal shadowing"). The gray values of the image correspond to the first derivative of the surface function of the object, so that the original function, the surface relief, can be generated by mathematical integration. However, surface relief reconstructions cannot provide the intrinsic 3D structure of objects – this is the domain of tomographic approaches.

Two special metal evaporation approaches, that is, rotational "shadowing" and decoration, are particularly suited to identifying regular structures of macromolecular complexes and fibrillar assemblies (e.g., of actin filaments or flagella). Decoration effects occur if only a limited amount of metal is evaporated so that there is no coherent metal coat. Even if the object

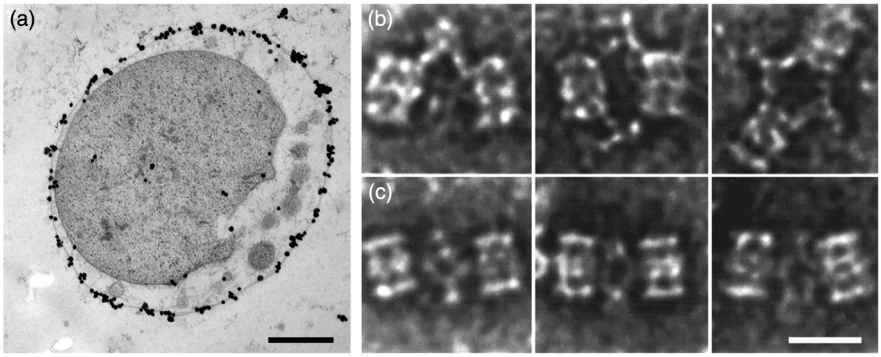


Figure 19.7 Labeling of proteins with antibodies. (a) Localization of the ATP synthase A1 subcomplex in an ultrathin section (after freeze substitution and embedding in Epon) of *Ignicoccus hospitalis*. The primary antibodies are labeled by secondary ones carrying gold clusters and having been enlarged by silver coating to 30–60 nm in diameter. The antibodies identify the enzyme in the outer membrane of the archaeon. (b) Immunolabeling of α- and (c) β-subunits of the negatively stained 20S proteasome from *Thermoplasma acidophilum*. The antibodies are clearly visible and identify the α-subunits in the outer and the β-subunits in the inner rings of the enzyme. Scale bars indicate 0.5 µm (a) and 20 nm (b) and (c). Source: part (a) courtesy of Reinhard Rachel, Regensburg, Germany.

temperature is below −100 °C the evaporated metal clusters can still diffuse and reach preferred locations on the surface. These sites are now decorated and the regular arrangement is clearly detectable. To obtain pure decoration but no shadowing effects it is advisable to deposit the metal at an angle of 90° onto the surface. Noble metals (Ag, Au, Pt) are particularly appropriate for decoration and often find different molecular sites.

Immunological Techniques, Chapter 5

Metal coating of isolated macromolecules is more complicated than negative staining and less rewarding than electron microscopy of native and vitrified specimens.

19.2.4 Labeling of Proteins

Negative staining and metal coating are non-selective contrasting methods. To identify and localize proteins amongst a wealth of other macromolecules we need specific labels such as monoclonal or polyclonal antibodies or other specific compounds that are coupled to gold clusters to achieve contrast in electron micrographs.

Thin sections, and particularly cryosections prepared for immunological purposes (method according to K.T. Tokuyasu), may be labeled with secondary antibodies bearing gold clusters (5–20 nm) that bind to the primary antigen-specific antibody and identify its position with about 20–30 nm accuracy (distance between gold cluster and antigen; Figure 19.7). Subsequent silver coating enhances the size and contrast of small gold clusters. Antibody labeling of isolated macromolecules does not require secondary enhancement by gold since the target protein as well as the antibody – and thus the specific contact regions – are visible with much higher spatial accuracy. Such experiments are particularly suited for identifying the position of subunits in heterooligomeric protein complexes (Figure 19.7).

19.3 Imaging Process in the Electron Microscope

The appropriate interpretation of images of a biological object does not only depend on its preparation and contrasting history but also on the imaging process in the electron microscope. It is thus useful to get some insight into the principles of image formation. It is sufficient to discuss basic physical concepts here; the comprehensive theory is described in specific textbooks.

19.3.1 Resolution of a Transmission Electron Microscope

Louis de Broglie (1892–1987) introduced the relationship between the wavelength λ and the momentum p of a moving mass. His equation can be transformed for electrons with nearly relativistic speed c (in the EM about 200 000 km s^{-1}):

$$\lambda = hc\frac{1}{\sqrt{2E_0E + E^2}}$$
$$E = U_b e_0 \qquad (19.1)$$
$$\lambda \approx 3.7U_b^{-0.6}$$

where Planck's quantum of action h is 6.63×10^{-34} J s, E_0 is the rest energy (8.19×10^{-14} J), and E is the kinetic energy of electrons. The latter depends on the accelerating voltage U_b (eV) in the microscope and the elementary charge e_0 (1.60×10^{-19} C) as given in Equation 19.1. The wavelength of electrons above ≈50 000 eV can be assessed by empirical approximation (given in nm). The wavelength amounts to 0.0037 nm for 100 000 eV electrons and is thus smaller than the wavelength of visible light by a factor of $\approx 10^5$.

Ernst Abbe attributed the resolution limit d of the light microscope to the physical parameters in Equation 19.2, which is also valid for the EM:

$$d = \frac{\lambda}{2n\sin\alpha} \tag{19.2}$$

Here the refractive index n is about 1, and α denotes half of the aperture angle (beam width) of the objective lens. This angle identifies the region of the diffracted beam that the objective lens can capture. Only these electrons transmit information about the object; those that do not interact with the specimen are blind to the object and belong to the reference beam. To obtain images with visible structures of size d (Abbe used a periodic pattern with the characteristic distance d) the lens must at least record a beam of the first diffraction order. Since the diffraction angle Θ is related to the inverse of d ($\Theta \approx \lambda/d$), the limiting angle α for the objective should be as large as possible ($\alpha \to \pi/2$). The term $A = n \sin\alpha$, that is, the numeric aperture, is of the order of 1 for objectives in light microscopes, but only ≈0.01 for electron microscopes.

Resolution Hermann von Helmholtz (1821–1894) introduced an alternative analysis of the imaging process in optical instruments and derived another definition of resolution. The beam, originating from a luminescent object, is diffracted by a lens of finite size that creates a diffraction disc (Airy disc) in the focus plane instead of a distinct diffraction spot. Two object points are still distinguishable if the diffraction maximum of one point falls into the first minimum of the second one (criterion of John Rayleigh, 1842–1919). Using this approach for a microscope with aperture $A = n \sin\alpha$ and replacing the viewing angle by the corresponding distance d, we obtain the formula $d = 0.61(\lambda/(n \sin\alpha))$. The resolution limits according to Abbe and von Helmholtz differ slightly by a numerical factor, and we also have to take into consideration that the nominator in Equation 19.2 depends on the illumination of the object, that is, parallel or oblique, and can thus achieve values between A and $2A$. The determining variables for resolution are wavelength and the numerical aperture. Using a small aperture that increases the image contrast leads to reduced resolution; the microscopist apparently sees better but recognizes less.

Owing to the small wavelength and despite the drawback of a small aperture, electron microscopes have a physical resolution of ≈0.2 nm with accelerating voltages of 100–300 keV. High voltage transmission electron microscopes (HVTEMs), equipment that is used in material science, are operated at voltages up to 10^6 eV. The corresponding wavelength of electrons (≈0.9 pm) thus allows for the resolution of single atoms.

19.3.2 Interactions of the Electron Beam with the Object

Imaging and Information Content of Diffracted Electrons in the TEM We distinguish between two kinds of interactions that electrons experience while passing through the object. The two kinds add to the imaging process in different ways and can be separated in certain microscope models.

The electrostatic interaction of a beam electron with an atomic nucleus of the object results in a deflection of the electron path. The deflection is strong if the negatively charged electron comes close to the positively charged nucleus, if the nucleus charge is high (high atomic number elements, staining with heavy metal) and if the velocity of the electron is low (the velocity is a function of the accelerating voltage). The energy of the beam electrons remains constant if the deflection angle is not too high and the electrons experience elastic scattering. The mean scattering angle is about 0.1 rad (≈6°). Part of the electron beam hits the aperture, is excluded from the imaging process, and creates the scattering contrast (Figure 19.8). The drastic suppression of scattered electrons thus creates high contrast in the image but limits the aperture angle and reduces the resolution of structural details (Equation 19.2).

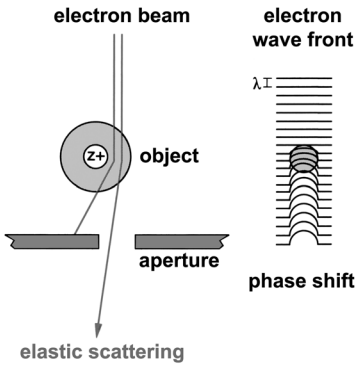


Figure 19.8 Interaction of beam electrons with the nucleus of object atoms. Strongly scattered electrons are shielded by the aperture and give rise to the scattering contrast. Electrons interacting with the object are decelerated by the object potential and show a phase shift of the wave front with respect to the reference beam. Interference of the scattered electrons with the non-interacting reference beam creates the intrinsic phase contrast in the transmission electron microscope (caused by lens aberrations). The scanning transmission electron microscope (STEM) is equipped with an electron detector instead of the objective aperture and records the strongly scattered electrons for imaging. The signal can be used for electron microscopical mass determination of macromolecular structures.

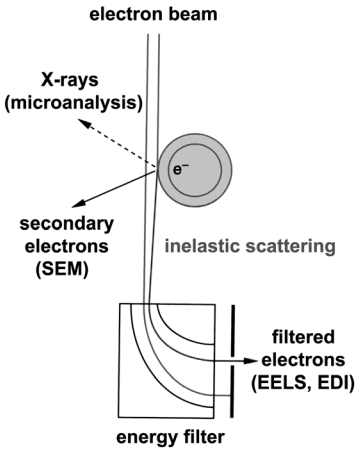


Figure 19.9 Interaction of beam electrons with the electron shell of object atoms. The beam electrons transfer energy to object electrons and this energy loss causes an increase of the wavelength. The beam electrons can be separated according to their energy in an energy filter and used for elemental mapping and analysis. The energy transfer leads to an excitation of object electrons and the emission of X-rays that are also indicative for the elemental composition of the sample. Emitted (secondary) electrons from (and beam electrons reflected by) the object surface are employed for imaging in scanning electron microscopy.

Since the interaction between beam electrons and the object is relatively strong, it is necessary to limit the thickness – or the mass density – of the specimen. This is generally the case with protein complexes, biological membranes, and ultrathin sections of embedded cells. In cryo-electron microscopy voltages of ≈200 keV are preferred for the examination of isolated macromolecules since they are usually embedded in a thicker layer of ice. Bigger objects such as intact cells (thickness ≥ 0.5 µm) in cryo-electron tomography (Section 19.5.3) can only be imaged at higher voltages (≥300 keV) without massive scattering and shielding of electrons.

Interactions of the beam with the electrons of the object have multiple, partly undesirable consequences for imaging. One effect is that the beam electrons are also scattered by Coulomb forces, but the typical scattering angle of $\approx 10^{-5}$ rad is much smaller than with electron–nucleus interactions. When accelerated electrons hit the electron shell of atoms they lose energy. The kinetic energy is reduced by ΔE and the wavelength increases correspondingly (Equation 19.1; Figure 19.9); this effect is termed inelastic scattering. The formerly almost coherent electron beam becomes incoherent and shows a spectrum of wavelengths after object transition. Since the diffraction depends on the wavelength ($\Theta = \lambda/d$) the optical system produces differently sized projections of the object and superposes them in the final image (which is equivalent to chromatic aberration of glass lenses). The structures become blurred and reduced in contrast, which is particularly problematic for objects with a high mass density, such as large frozen-hydrated macromolecular complexes, viruses, cell organelles, or intact cells. In general, thick ice layers considerably increase the proportion of inelastically scattered electrons. It is, however, possible to separate electrons of lower energy from the elastically scattered ones by means of an energy filter (Section 19.3.5).

Electron Energy Loss Spectroscopy The energy loss of beam electrons ($\Delta E \leq 2000$ eV) correlates with the energy uptake by object atoms and thus contains information on the interacting elements. The spectral analysis of inelastically scattered electrons reveals the elementary composition of the object, and since these electrons can also be used for imaging we can record images of the element distribution (elementary map). Spectra are obtained by electron energy loss spectroscopy (EELS) and images by electron spectroscopic imaging (ESI) or electron spectroscopic diffraction (EDI). For this purpose the microscope must be equipped with a magnetic prism that sorts electrons according to their energy and filters those that are not required for imaging. EELS is usually applied to thin sections of cells or tissues; it is not suitable for single molecules and vitrified specimens.

Mass Determination in the Scanning Transmission Electron Microscope The strongly elastically scattered electrons that are shielded by the object aperture also contain information about the biological object since they were deflected according to the number of protons in the nucleus and the amount of corresponding atoms in the object. These electrons can be utilized for imaging if we replace the aperture with an electron detector (Figure 19.8) in the scanning transmission electron microscope (STEM). Provided the elementary composition of the biological specimen is known, which is the case for proteins to a good approximation, we can interpret the signal intensity as a measure for the mass of the object and determine the molecular mass of protein complexes and other biological structures. The sample must not be embedded in any other material and is investigated in a native, freeze-dried state. It is sufficient to record the signal of several hundred (or more) particles for a reasonable statistical analysis with an accuracy of about ±5%. STEM mass determination of isolated proteins is thus not superior to mass spectroscopy.

Mass Spectrometry, Chapter 15

However, the method does not depend on the size and structure of the material and it is possible to determine the mass of large and arbitrarily formed heterooligomeric and multimeric complexes of macromolecules. Typical tasks for STEM mass determination are the stoichiometric analysis of protein complexes, the mass per length of fibrillar structures such as flagella and microtubules, and the mass per area or per structural feature of membranes, 2D crystals, or other 2D assemblies. Mass mapping is only possible with the STEM technique.

Scanning Electron Microscopy and Analytical Electron Microscopy Any interaction of electrons between beam and object retains the energy and momentum of the entire system. This means that the energy loss of beam electrons increases the energy of object electrons. They are excited and occasionally even expelled from the electron shell. This process is accompanied by

emission of electromagnetic radiation, heat, and electrons from the irradiated material. These secondary electrons are used for imaging in the SEM. The electrons originate from the surface of the object and create images of cells and small organisms with impressive depth of sharpness. The resolution power of the conventional SEM is limited (≥10 nm) and is usually too low for imaging protein complexes. However, low voltage scanning electron microscopes (LVSEMs) are able to record small structures.

Electrons having been removed from the electron shell leave behind a gap that is filled by electrons from a higher energy level. The energy difference is emitted by radiation where electron gaps in the K shell entail particularly energy-rich radiation, that is, X-rays. These contain information on the elementary composition of the material, similar to the EELS spectra of the beam electrons. Energy dispersive X-ray spectrometers (EDSs) detect signals from elements above the atomic number 10 (EELS spectra are sensitive down to atomic number 4). The method of X-ray microanalysis is usually applied with ultrathin sections and inorganic samples.

19.3.3 Phase Contrast in Transmission Electron Microscopy

Thin biological specimens that mainly consist of elements of low atomic number (H, C, N, O) behave as weak phase objects in the microscope. The object potential (equivalent to the refractive index in light microscopy) decelerates scattered electrons, and the electron wave experiences a small phase shift $\Delta\phi$ with respect to the non-interacting reference beam when leaving the object (Figure 19.8). The intensity remains (almost) unmodified. Our eye, digital cameras and photo-emulsions cannot record phase differences and the object would actually be invisible. Looking at the difference between the weakly phase-shifted and the reference wave, we obtain a wave of identical wavelength, smaller amplitude, and a phase shift of about $\pi/2$ or $\lambda/4$ with respect to the reference. If it were possible to shift the reference wave by $\pi/2$ too, so that the amplitudes would be aligned, the waves would interfere and produce a detectable amplitude modulation. In light microscopy, the diffracted and the reference beams are conducted through a glass plate (phase plate) that is of different thickness for the reference beam (the beams are separated from each other in the back focal plane where the phase plate is located) and creates the desired phase shift. The non-diffracted beam is, moreover, attenuated, which enhances the amplitude modulation, that is, the phase contrast.

The situation is more complicated in the TEM. Here, it is sufficient to mention that the spherical aberration of the objective lens (characterized by the parameter C_S) generates phase contrast as a function of the scattering angle and that it can be adjusted by varying the focus. The ideal adjustment is a weak underfocus, known as the Scherzer focus after Otto Scherzer (1909–1982) who carried out the theoretical calculations. However, the phase shift and thus the contrast, is neither constant nor ideal over the complete scattering range, and it is particularly low with Scherzer focus conditions. The focus-dependent contrast transfer function (CTF) describes the contrast contributions for structural details of size d (corresponding to spatial frequencies related to $1/d$; Section 19.4.3). The phase contrast may be strong or weak, become zero, or even change its sign. EM images contain more or less completely or correctly transmitted object information as a function of focus. Thus, the appropriate interpretation of images always requires analysis and ideally the correction of the CTF (Section 19.4.3).

19.3.4 Electron Microscopy with a Phase Plate

If there was a phase plate for electron microscopes analogous to the one in light microscopy, one could adjust ideal focus conditions at maximum contrast. The physicist Hans Boersch (1909–1986) had already introduced the basic ideas in 1947, but its realization was hampered by technical problems. Only recently have technical developments changed the situation.

The basic principle is to apply an electric potential to the reference beam (or to the scattered one) to create the required phase shift of $\pi/2$. This is obtained by an electrostatic potential in the center of the back focal plane for the unscattered electrons (Boersch phase plate) or by a thin carbon film with a small central hole, leaving the reference beam unchanged (Zernike type phase plate; Figure 19.10a). While both solutions and variants thereof were realized, they did not find their way into everyday applications. In particular, biological EM has not used

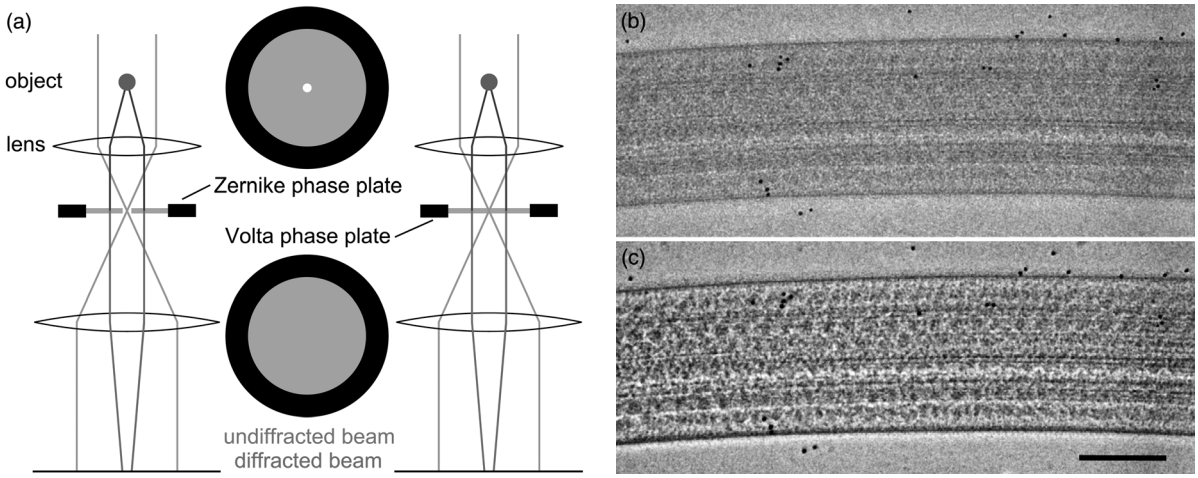


Figure 19.10 Phase contrast by phase plates in transmission electron microscopy. (a) Scheme of the optical path in the microscope. The undiffracted (zero) beam is focused in the back focal plane below the object lens. The Zernike type phase plate possesses a central hole so that the zero beam can pass through whereas the diffracted beam crosses the phase plate material (carbon film) and experiences a phase shift due to the positive inner potential of the material. The ideal result is a phase shift of $\pi/2$ compared to the zero beam (positive phase contrast). The Volta phase plate is a continuous (carbon) film heated to >200 °C to avoid contamination during irradiation. The focused zero beam induces a negative surface (vacuum) potential (Volta potential) that overcompensates for the positive inner potential of the material and effectively shifts the phase by ideally $\pi/2$. The object is imaged with positive phase contrast. (b) Electron microscopic projection of part of a native frozen-hydrated and unstained worm sperm taken without a phase plate and (c) with a Volta phase plate illustrating the increase of contrast. The black dots are gold markers. Bar indicates 200 nm. Source: courtesy of Maryam Khoshouei, Martinsried, Germany.

electrostatic phase plates, and the thin film Zernike-type phase plate suffers from charging and produces fringing in images, an unavoidable diffraction effect of the sharp edge of the central hole. The effect can be lessened by image processing, but it still affects the image quality. A novel type of thin film phase plate, the Volta potential phase plate, exploits the electric (surface or vacuum) potential created by the high intensity of the reference beam in a contamination-free carbon film (Figure 19.10). This principle has only recently been discovered, and the formation of the electric potential is not yet fully understood. However, this tool avoids fringing (there is no hole), contamination (by heating), is stable, reusable, and does not require complicated alignments – an important advantage for routine and frequent applications in cryo-EM and cryo-electron tomography. The resulting phase contrast is remarkable, so that this technique will likely become a standard application in biological EM of unstained and frozen material (Figure 19.10b).

19.3.5 Imaging Procedure for Frozen-Hydrated Specimens

Cryo-electron microscopy is the most important development in microscopic structure research of biological material. The technique enables us to examine molecules and cellular components at high resolution in a close-to-native state. However, chemically untreated and unstained samples are very radiation-sensitive and are rapidly destroyed upon irradiation. The electrons create molecular radicals that readily react with other molecules and cleave chemical bonds, leading to mass loss and eventually to structural destruction. The low temperature of liquid nitrogen (−196 °C) lessens the reaction rate and renders the biological object six times more resistant to radiation than at room temperature. Cooling with liquid helium (≈4 K) promises an even higher cryoprotection factor but unfortunately presents us with problems created by the loss of friction in frozen hydrated specimens.

Resolution-limiting radiation damage is indicated by bubble formation in the frozen sample (Figure 19.11). To avoid this effect the total electron dose should not exceed $\approx 100\,e^{-}\,\text{Å}^{-2}$, whether it is applied to a single projection or to a complete tilt series for 3D reconstruction (Section 19.5). In the latter case, the tolerable electron dose must be shared by all projections and adjustment procedures. Although it is technically possible to keep the dose arbitrarily low for each projection and to protect the specimen from any damage, one needs a sufficiently high

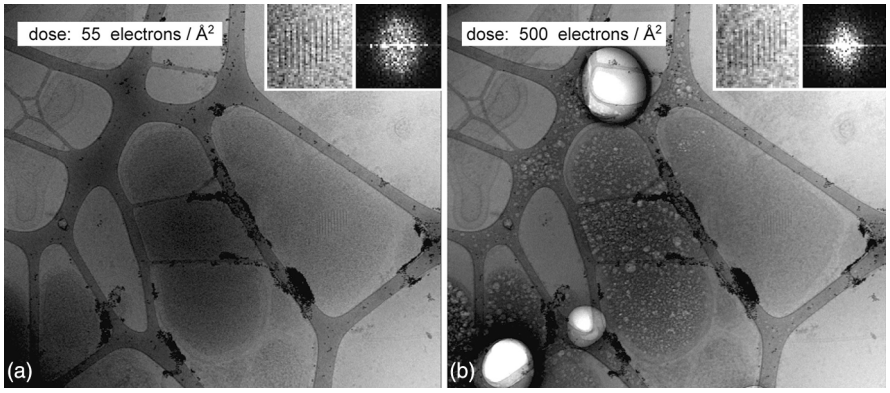


Figure 19.11 Cryo-electron microscopy of vitrified cells of the archaeon *Pyrodictium abyssi* imaged with (a) low and (b) high cumulative electron dose. The cells are embedded in ice between the bars of the carbon support. One of the cells contains a protein crystal (enlarged and displayed together with its power spectrum (PS) in the inset). The heavily irradiated image shows clearly attenuated spots in the PS and bubbles (bright) within and outside of the cell. They indicate massive beam damage of the biological material and ice. The cumulative dose must therefore be kept below a critical threshold to prevent detectable beam damage in the object. Source: courtesy of Stephan Nickell, Martinsried, Germany.

signal-to-noise ratio for subsequent image analysis and processing (Section 19.5.3). The only way between Scylla and Charybdis here is to use most of the electron dose for data recording and do all the adjustments at other object sites. In cryo-electron microscopy this process is automated.

The microscopist avoids direct inspection of the object on the fluorescence screen and instead records images by means of a camera. The appropriate procedure is to search the area of interest at low magnification, recording a test image with the desired magnification close to the object site, and determining the focus and other parameters offline. The program calculates the actual values for the object site and adjusts the imaging parameters of the microscope accordingly. The micrograph is recorded automatically, and the electron beam is deflected immediately after in order to minimize irradiation. In cases that require many (hundreds to thousands) images of single particles, the program selects a new site and thereby scans large areas of the grid. If using different projections from a single object, the program actuates the sample holder, turns it by a given angle, centers the object site again, adjusts the focus conditions, records the image, and continues accordingly until the series of tilt angles (tilt series) is complete. In this way, almost the entire electron dose is available for image recording.

Projections of thick, ice-embedded samples contain a significant amount of information originating from inelastically scattered electrons that attenuates the image contrast considerably (Section 19.3.2).

The use of energy filters in the mode of zero-loss filtering excludes undesirable electrons that experienced energy loss. Only elastically scattered ones can pass, resulting in improved image quality (Figure 19.12). The contrast enhancement is physically independent of the contrast created by phase plates. The techniques are complementary and they both contribute to optimizing image contrast and quality.

Energy filter Electrons of various energy states differ in their frequency and wavelength ("color"). They are individually diffracted by the electron lenses and project into slightly different positions in the final image. The projected structures are thus of varying size and attenuate the sharpness and contrast upon superposition in the final image. Electromagnetic energy filters widen the electron beam according to the spectrum of electron energies, and an adjustable aperture selects "monochromatic" electrons of identical wavelength for imaging or analysis. Energy filters correspond to color filters in optical microscopy.

19.3.6 Recording Images – Cameras and the Impact of Electrons

The last step in image formation is the detection of the imaging electrons. In recent years, digital cameras have replaced film plates, and they are now the usual medium for recording electron micrographs. CCD (charged coupled device) cameras record electrons in a scintillator layer that emits photons instead. By repeating this process, these photons produce a shower of new photons and thus increase the signal while also broadening it at the expense of signal intensity of high-resolution information. The camera imprints another characteristic on the recorded image, namely, the modulation transfer function (MTF), which modulates (attenuates) the signal of structures according to its spatial frequency. Recently, the construction of direct electron detectors yielded two important advantages. Firstly, the incoming electron is recorded in the respective pixel of the detector and not spread over a couple of neighboring pixels, that is, the spatial resolution is high. The MTF is considerably increased for higher spatial frequencies, and small structural details can be detected much better. Secondly, the detector is very fast and allows images to be read out in milliseconds, much faster than with CCD cameras. Recording movies (a number of frames) instead of one single (final) image revealed that the objects are moving by beam-induced effects. These movements blur images taken by conventional

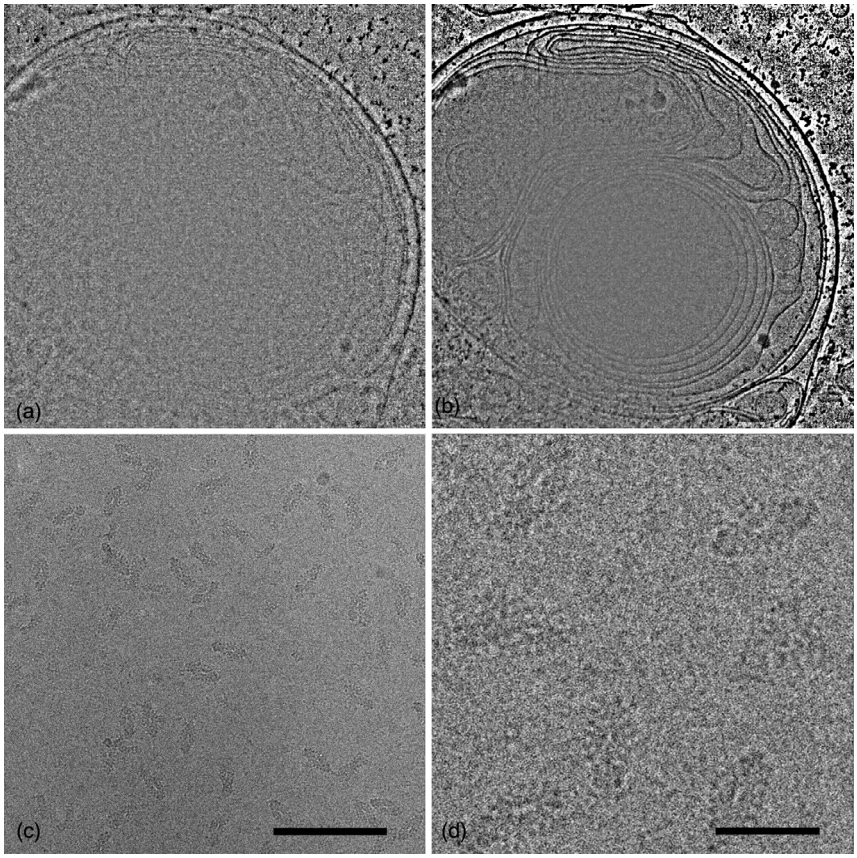


Figure 19.12 TEM images of vitrified lipid vesicles ((a) and (b)) and of frozen-hydrated enzyme complexes tripeptidyl peptidase II from *Drosophila melanogaster*, (c) and (d). (a) Electron micrograph without energy filtering; elastically and inelastically scattered electrons contribute to the image. (b) Only elastically scattered electrons formed the phase contrast image. Inelastically scattered electrons with lower energy were filtered out. The multiply nested vesicles are now clearly visible. (c) The enzyme complexes are poorly detectable in the original micrograph because of the low contrast and signal-to-noise ratio. (d) The macromolecules in different orientations with higher magnification. Classification of equivalent projections are shown in Figure 19.19. Source: part (b) courtesy of Rudo Grimm, Martinsried, Germany; part (d) courtesy of Beate Rockel, Martinsried, Germany.

cameras and also destroy high-resolution information. But the shifts can now be corrected by aligning the frames of a series and adding the frames to a well-resolved and unblurred image. The new detectors push the quality of appropriate single-particle reconstructions to quasi-atomic resolution, a "quantum step" in electron microscopic imaging.

The first step in electron microscopy – recording images of the biological sample – is now complete. The second part deals with data analysis and image processing for 3D reconstruction.

19.4 Image Analysis and Processing of Electron Micrographs

Electron micrographs contain the recordable signal of the object, although obliterated by unstructured noise and modulated by the contrast transfer function of the microscope and the signal transfer characteristics of the camera. The smaller the objects and the higher the expectations for the resolution of molecular details are, the more the contributions of noise and hardware effects affect the visibility of structures. Noise and effects of transfer functions must thus be determined and ideally separated from the desired signal. This is the business of image analysis and the processing of electron micrographs.

19.4.1 Pixel Size

Digital EM images are usually 2048 × 2048 ("2k"), 4096 × 4096 ("4k"), or more pixels in size. The beneficial primary magnification of the microscope depends on the pixel size of the camera and the desired resolution of the image. Harry Nyquist (1889–1976), Claude E. Shannon (1916–2001), and others described the minimal condition for the resolution of structural details

in digital images. A structural element can be regarded as being resolved if it is defined by at least two pixels (in one dimension). If P_c is the pixel size of the camera and M the primary magnification of the microscope (including corrections for the optical distance to the camera) the pixel size on the object level is $P_o = P_c/M$ and the resolution limit $d = 2P_o$. The pixel should not exceed ⅓ or ¼ of the desired resolution to compensate for some loss upon interpolations (Equation 19.3) and to cope with the limited sensitivity of cameras (Section 19.3.6):

$$P_o = \frac{P_c}{M} \leq \frac{d}{3} \tag{19.3}$$

The visibility of tiny structures depends on the physical resolution of the microscope (Equation 19.2), on the parameters of image recording, the type and quality of the camera, and on the delicate preparation of the biological object (fixation, staining). The shape of negatively stained protein molecules can be resolved to ≈2 nm. Parts of this size consist of about 30 amino acids (N_{AA}) with a mass of ≈3400 Da (M_{PROT}). These values, as well as the domain volume V_d (nm^3), can be assessed with Equation 19.4. The factors 3.9 or 7.6 (nm^{-3}) and 430 or 840 (Da nm^{-3}) derive from the protein density (≈1.41 g cm^{-3}) and the average mass per amino acid (≈110 Da) in proteins. The average protein density depends on the protein size and increases to about 1.5 g cm^{-3} as the protein size changes from 30 to 10 kDa. Often, the molecular water layer on proteins is part of the volume determinations and reduces the average density to 1.37 g cm^{-3}, a value that is commonly applied in calculations:

$$\begin{aligned} N_{AS} &\approx 3.9d^3 \approx 7.6V_d \\ M_{PROT} &\approx 430d^3 \approx 840V_d \end{aligned} \tag{19.4}$$

Vitrified proteins yield structural information to ≈0.5 nm or even better after image processing. Thus, the secondary structure and in ideal cases even the density of amino acid residues is resolved. Cameras with a typical pixel size of 15–20 µm suggest a primary magnification M of 30 000–100 000. Higher magnifications are usually not necessary and would increase the radiation burden, which is related to M^2.

19.4.2 Fourier Transformation

Many operations for image analysis and processing are not performed with real data but with its Fourier transform (FT). It is used, for example, to judge the quality of the contrast transfer function (CTF) and to analyze structural regularities of objects. The FT is also required for improving the signal-to-noise conditions in low contrast images and for certain steps in averaging and 3D reconstruction.

It is helpful to know what characteristics a Fourier-transformed image has, but it is not necessary here to introduce the mathematics developed by Joseph Fourier (1768–1830). Let us consider a one-dimensional image, that is, a single pixel line of a micrograph, representing the intensity curve of an object. Fourier's insight was that almost all continuous curves can be represented by the sum of (infinitely many) basic sine functions of varying frequencies (Figure 19.13). To distinguish the sine oscillations along a distance in space (e.g., along the x-direction of an image) from the oscillations in time (frequency) we use the term *spatial frequency*. The inverse or reciprocal of the spatial frequency, the wavelength, identifies a certain distance d in the real image, and the amplitude of the sine wave is related to the density oscillation along d. Mathematically, the sine function has the amplitude 1. To adjust it to the corresponding density value in the image we need to multiply it by a factor. These factors describe the density distribution (i.e., the gray values) of the object's structure and are known as structure factors. Sine functions with small spatial frequencies (large wavelengths or values of d) belong to large object structures, and functions with high spatial frequencies (small values of d) to small structural details. Obviously, the macrostructure (e.g., the body of a hedgehog) is characterized by large structure factors and the fine-structure (e.g., the stingers) by much smaller amplitudes of the corresponding sine functions. If an image is to be analyzed to maximum resolution the very small structure factors have to be separated from contributions of the superposed high-frequency noise.

The amplitudes of sine waves are, however, not sufficient to exactly describe the curve of a structural density. Moreover, we have to know the exact location of the origin of each sine wave. The position of the corresponding density curves defines where the object (the hedgehog) is located and the substructures (the stingers) are arranged. Since the trigonometric function is

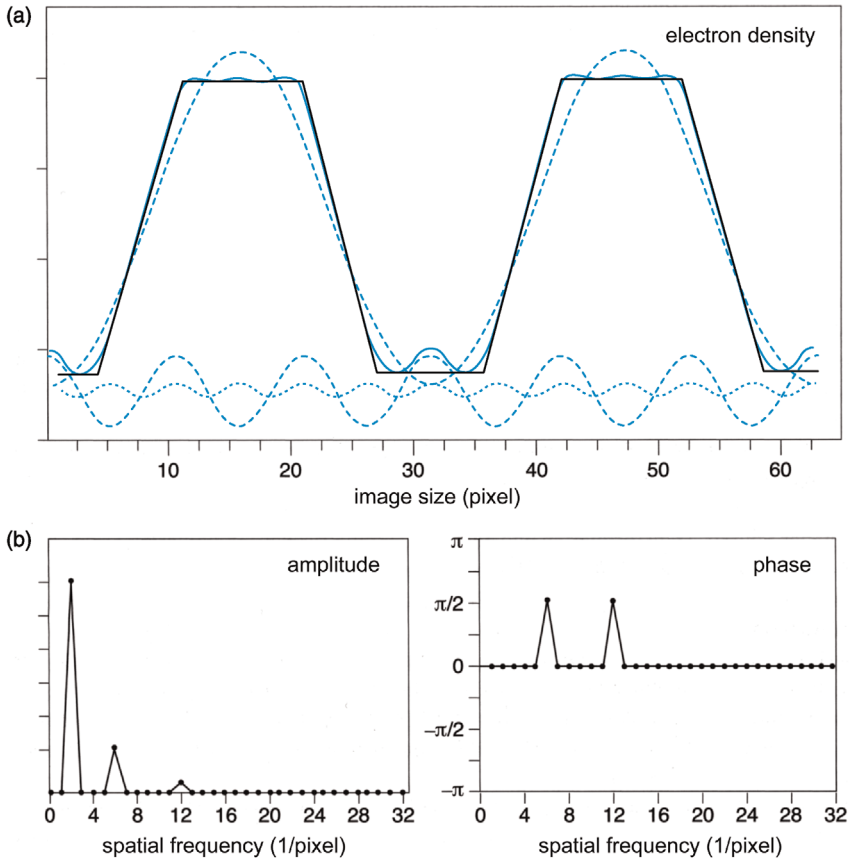


Figure 19.13 Fourier analysis (Fourier transformation) of a one-dimensional object (a, curve in black). The superposition of three sine functions (broken lines in red) approximately traces the original structure (continuous line in blue). The object could be exactly reconstructed by (infinitely) many sine functions with increasing spatial frequencies, appropriate amplitudes, and phase shifts. Plotting the amplitudes and phase shifts of all constructive sine functions against the spatial frequency produces the Fourier transform of the original image (b). The transform has basically the same size as the original image, but it contains equivalent Fourier data symmetrically to the center that are related according to Friedel's law (Georg Friedel, 1865–1933). This relation is clearly defined and so it is sufficient to show only one half of the Fourier transform.

periodic, the largest difference between the origin of a particular sine wave and a common reference point (e.g., the origin of a coordinate system in the center of an image) is the wavelength of the corresponding sine function. This deviation is called the phase shift ($-\pi \leq \Delta\phi \leq \pi$), often (but incorrectly) referred to as *phase*.

The object function is thus fully described by the sum of elementary sine waves with (continuously) increasing frequency ν_i and the corresponding individual amplitudes k_i and phase shifts $\Delta\phi_i$. A description of an image with an infinite number of sine functions is impractical. Since we usually record digital images of n pixels in size (e.g., in the x-direction) we subdivide the object function F into n pixels with $n/2$ spatial frequencies, that is, basic sine waves. The smallest frequency ν possesses a period of exactly 1 with a wavelength $\lambda = n$ pixels corresponding to the image size S ($S = n$ pixels where $\nu = S/\lambda$). The highest frequency (Nyquist frequency) is characterized by the smallest possible wavelength with $\lambda = 2$ pixels. The object function F is the sum of all possible (finite) sine functions (Equation 19.5). The pixel number s indicates a position in the image of size S, and x denotes the relative position ($x = s/S$). The value $F(x)$ describes the density (gray value) of a pixel at curve position x and is calculated by:

$$F(x) = \sum k_i \times \sin(2\pi\nu_i x + \Delta\phi_i) \qquad x = \frac{s}{S} \tag{19.5}$$

The Fourier analysis of an image is nothing other than the calculation of all the structure factors k_i and phase shifts $\Delta\phi_i$ that are necessary to scale the sine waves with frequencies ν_i and to place them with respect to the origin of the coordinate system (the image). Fourier transforms are therefore data ("images") that consists of two parts. One contains the amplitudes (structure factors), the other the phase shifts of the sine functions plotted against the spatial frequency ν. Since ν is related to the reciprocal of distance in real space, Fourier transforms are representations of images in reciprocal (Fourier or frequency) space. The Fourier transform $F(x,y)$ of a two-dimensional image requires two-dimensional sine functions (x,y) of course, but they are essentially operated the same way (Figure 19.14).

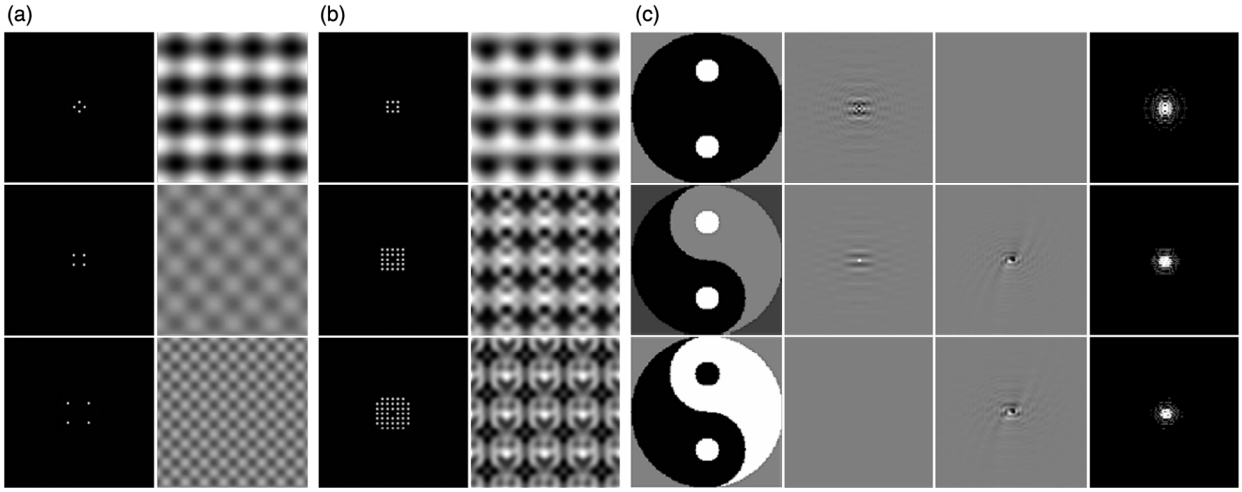


Figure 19.14 (a) Examples of two-dimensional sine functions (organized in a 2D lattice) with lower and higher spatial frequencies, varying orientations, and different amplitudes. The right-hand column shows the real images of the functions (bright areas indicate positive, dark areas negative values); the left-hand column shows the corresponding power spectra (PS). The spots in the PS characterize the spatial frequencies and the orientations of the lattices in the *x*–*y*-direction. (b) Examples of the synthesis of images through superposition of basic sine waves. The upper image contains the information from the first two images in (a), the other images increasingly more sine functions with different orientations and higher spatial frequencies. (c) Examples of symmetric and antisymmetric (non-periodic) images (column on the left), their Fourier transforms (FT, central columns), and power spectra (column on the right). The FTs consist of a symmetric (real) and an antisymmetric (imaginary) part each (see Equation 19.6). The Fourier data represent negative (dark) and positive values (bright); zero corresponds to a medium gray. The FT of the symmetric image possesses Fourier coefficients unequal to zero only in the real part, the FT of the antisymmetric Yin Yang symbol only in the imaginary part. The image in the middle contains symmetric and antisymmetric features that can be separated by setting one or the other of the two FT parts to zero prior to back-transformation.

Usually, the FT is given in an alternative form, which is obtained by a simple trigonometric operation (Equation 19.6):

$$k_i \sin(2\pi\nu_i x + \Delta\phi_i) = a_i \cos(2\pi\nu_i x) + b_i \sin(2\pi\nu_i x)$$
$$a_i = k_i \sin(\Delta\phi_i) \text{ and } b_i = k_i \cos(\Delta\phi_i) \tag{19.6}$$

The amplitudes and phase shifts can be calculated from the factors a_i and b_i. The FT of an image is again a two-part representation of the factors a_i and b_i as functions of spatial frequency. This representation has the advantage of possessing analytical properties. Since the cosine is a symmetric function with respect to the origin while the sine is an antisymmetric one, the part of the FT containing the factors a_i characterizes the symmetric features and the other part with factors b_i the antisymmetric features of the structure. The Fourier transform is a complex mathematical function (and usually written in an exponential form); the symmetric part is thus also named the real part and the antisymmetric one the imaginary part of the FT (Figure 19.14). Image processing systems make use of the fast Fourier transform (FFT), which is an efficient algorithm for transformation of digital (discrete) data of certain dimensions.

It is often sufficient to know the structure factors (intensity, power) and their distribution in reciprocal space to get some idea of object characteristics and image quality. These are provided by the power spectrum (PS), that is, the product of the FT with its complex conjugate FT*. The PS is the transform of the auto-correlation function of an image or the "cross-correlation" function of an image with itself (Section 19.4.4). The PS misses the phase information and contains the squared structure factors only. It corresponds to the light-optical diffractogram that is created by diffraction of coherent light (e.g., of a laser beam) passing through a micrograph of a structure (Figure 19.14). Calculations of Fourier transforms, power spectra, and correlation functions are standard operations in program system for the analysis, processing, and 3D reconstruction of images.

19.4.3 Analysis of the Contrast Transfer Function and Object Features

The electron microscopic image is a function of the projected object structure and the contrast transfer function of the microscope that depends on the electron optical characteristics and the

actual adjustments of imaging parameters. The object structure (the density function) is convolved with the transfer function during the imaging process. Mathematically, this means that the functions are multiplied in Fourier space, and this means they can be analyzed in Fourier transforms of images.

Contrast Transfer Function We already know that the diffraction of the electron beam by the object results in positive and negative phase contrast, depending on the diffraction angle, and that regions in-between are missing contrast, that is, the structural factors of corresponding spatial frequencies are (close to) zero (see Section 19.3.2). The transfer function is particularly clear in projections of a thin amorphous carbon film. Depending on the focus adjustments the power spectrum of such an image shows several bright rings separated by dark gaps. The bright regions represent the squares of the structure factors of transferred spatial frequencies. The dark rings denote gaps in information transfer (contrast). The intensity of corresponding spatial frequencies is very low, or even zero, and thus eliminated. This pattern is known as Thon rings (Figure 19.15).

The structure factors of the object that falls into these gaps are eliminated as well, and the corresponding information is thus missing in the electron micrograph. Structure details represented by spatial frequencies falling into the region just beyond the first gap are imaged with inverted contrast and so on. Such effects can hardly be identified by eye in projections of biological specimens, but a series of images taken with different focus values illustrates the effects (Figure 19.15). There is an ideal focus level (moderate defocus) that guarantees a continuous transfer of contrast to a spatial frequency of $\leq 1\ \mathrm{nm}^{-1}$ and that is advantageous for biological electron microscopy of (stained) protein complexes. This focus level is close to the absolute contrast minimum and it requires some training to prefer it to the allegedly clear, contrasty but heavily defocused adjustment.

If we try to obtain reconstructions with maximum resolution we have to correct for the contrast reversal of certain spatial frequencies afterwards. The Fourier coefficients between the first and the second gap (and third and fourth gap, etc.) are "flipped" by multiplication by -1, and images with different focus levels fill the missing information of the unavoidable gaps (producing different CTFs).

Two typical imaging aberrations, that is, astigmatism and drift, are easily identified in power spectra (and should encourage the microscopist to discard those data). Astigmatism is the effect of different focus states in perpendicular image directions. It stems from distortions of the electromagnetic field in the lenses of the microscope. The Thon rings become ellipses and hyperbolas close to in-focus ("Gauss focus") situations. The gapless transfer of the signal to a certain spatial frequency (casually referred to as resolution) may therefore change dramatically in different directions of the image. Such micrographs typically exhibit streak artifacts. Objects or specimen holders that drift while the micrograph is being recorded produce blurred images in the direction of movement and Thon rings with partially reduced intensities. It is an essential part of quality control to evaluate the transfer function and then select suitable electron micrographs for further processing.

Object Characteristics The Fourier transform of an amorphous structure, for example, of single protein molecules, is not very informative for the interested observer (Figure 19.14). However, the situation differs for regularly arrayed macromolecules in 2D lattices. Here, the same amplitudes and phase shifts occur as often as molecules exist in the 2D crystal, and they accumulate to a significant spot (reflex) at corresponding spatial frequencies in the PS. Moreover, the frequencies describing the regular structure must be related to each other, that is, the higher frequencies are always whole-number multiples of the lowest frequency since all the sine waves must have identical relative positions with respect to all molecules in the lattice. Sine functions that do not hit identical reference positions of each molecule with the same phase (and thus with the same intensity) are not suited to describe the structure of the crystal and are blanked out; their structure factors are zero. This is the case for most frequencies and only a few constructive ones are left, which are identified by a number of regularly arrayed diffraction spots in the PS (Figure 19.14). Irregularly distributed data in-between originate from other, non-crystalline contaminants and noise. The arrangement of diffraction spots yields information on the crystal structure such as the orientation and type of the lattice (i.e., tetragonal, hexagonal, and others), the lattice constant (the periodic distance between molecules in the crystal), and the characteristic angle between the lattice vectors

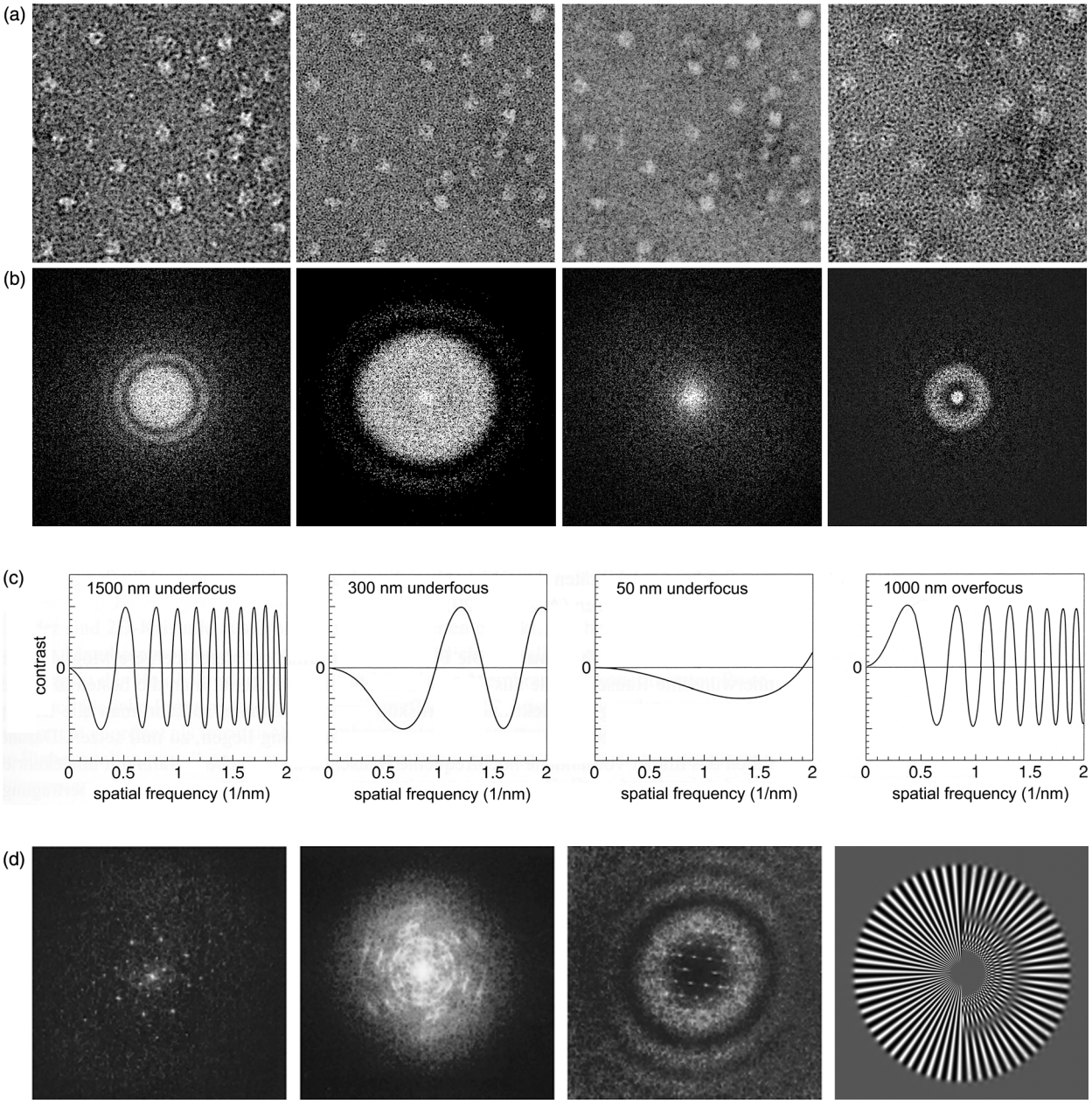


Figure 19.15 (a) Focus series of a negatively stained protein complex (diameter ≈12 nm). The images were taken with strong underfocus (maximum contrast), moderate underfocus, in focus condition (Gauss focus with minimum contrast), and strong overfocus. The grain pattern is particularly dominant in strong defocus conditions. (b) The corresponding power spectra illustrate the effect of the contrast transfer function that varies with focus. Dark gaps between the bright Thon rings denote missing structure information. These gaps already occur in regions of low spatial frequencies, that is, with relatively large object structures, in cases of strong defocus. The first gaps in the series are located at spatial frequencies of $(3.2\,\text{nm})^{-1}$, $(1.4\,\text{nm})^{-1}$, $(<0.8\,\text{nm})^{-1}$, and $(2.9\,\text{nm})^{-1}$. (c) One-dimensional contrast transfer functions (CTF) without signal damping. It is obvious that the contrast of low frequency information is inversed in the case of overfocus conditions. This is why the stain-filled central hole of the protein molecules (top right image in (a)) is now bright and no longer dark. (d) Power spectra of EM images of 2D protein crystals (images 1–3 from left to right) show reflections typical for a good crystal without or only with subcritical lattice distortions, a 2D crystal with massive lattice distortions and correspondingly unclear spots, and a good crystal recorded with strong underfocus. Some of the reflections fall into the transmission gaps and are eliminated, others are in regions with inverted contrast and destroy the structural information. The attenuated rings in the x^*-direction indicate drift of the specimen during image recording. The last image in (d) shows a resolution test pattern in the original form (left half) and after convolution with a contrast transfer function creating gaps of missing information (gray) and contrast reversals after each gap. The originally bright and dark rays now oscillate in their contrast depending on the spatial frequencies in the Fourier transform.

(e.g., 60° with a hexagonal crystal). We also obtain information on the lattice quality (minimal or massive lattice distortions) and the crystallographic resolution (Figure 19.15) that is indicated by the spatial frequency of the highest diffraction spot visible.

19.4.4 Improving the Signal-to-Noise Ratio

Electron micrographs contain two types of signals, the desired one that shows the structure of interest and noise that originates from other sources. Noise is not correlated with the structure of the biological object and varies from one image area to the next.

Adding up the gray values of many identically projected protein complexes increases the object signal and averages statistical noise to an approximately constant and structureless value. This process even enhances structural signals that cannot be distinguished from noise in single projections. The signal-to-noise ratio (S/N) increases with the square root of the number of averaged projections. If the initial S/N ratio is already high it is sufficient to average a few thousand particles. Frozen and unstained protein complexes require 10^4–10^5 or more single projections to achieve the maximum resolution. Apart from the effects of preparation and the number of particles, there is another resolution-determining factor, that is, the accuracy of particle alignment. The precision depends on the contrast, which is usually low in frozen-hydrated preparations if not enhanced by application of a phase plate. The second limitation arises from structural and conformational variabilities of protein complexes. They can be tackled by identifying structural classes consisting of coherent data sets for averaging and reconstruction (Section 19.4.5).

Filtering in Fourier Space Fourier transforms offer several possibilities to eliminate undesirable noise from images. If we know the resolution for our data we can set all spatial frequencies higher than the resolution limit to zero. The same operation can and should be performed with spatial frequencies that lie beyond the first gap of the contrast transfer function (if it is not corrected) in order to avoid corruption of structural details by the effects of inversed contrast (Figure 19.15). This low-pass filtering increases the contrast in the back-transformed image. High-pass filtering removes low frequency contributions such as large contrast variations in images of unevenly negatively stained preparations.

Filter Approaches for Crystal Data Images of ideal 2D crystals yield very sharp and far-reaching diffraction spots in Fourier transforms and power spectra that contain the complete structural information of the crystalline specimen. Spatial frequencies in-between do not contribute to the structure and can confidently be set to zero. This rigorous elimination of noise and other non-constructive contributions is termed window filtering since only small windows around the reflections are left open. Diffraction spots are usually spread over several pixels. To remove the remaining noise in these areas one may fit ideal profiles to the peaks and replace the original data. If the crystal unit cells possess rotational or mirror symmetry, interrelated diffraction spots must have identical amplitudes and certain phase shifts; these can be adjusted in a last step. The back-transformed image is now apparently noise-free and represents the Fourier synthesis of the structure. This procedure resembles data processing in X-ray crystallography. However, the diffractograms are equivalent to power spectra where the phase information is lost; it has to be recovered by separate experiments.

Unfortunately, 2D crystals of proteins are usually not perfect. High-quality crystals are obtained with membrane proteins, for example, the naturally crystalline bacteriorhodopsin from halobacteria, or other proteins that are reconstituted in artificial lipid membranes. Soluble proteins can be crystallized at interfaces, for example, at the interface of water and a lipid monolayer spread on the water surface. The crystalline order is usually not ideal and one has to cope with significant lattice distortions and orientation variations. Diffraction spots of higher spatial frequencies are now distributed over larger areas in Fourier transforms, are attenuated in intensity, and may even be hidden in reflections of noise.

Correlation Averaging Lateral dislocations of unit cells in imperfect 2D crystals may be corrected for by the individual determination of unit cells. To find these positions we need to cross-correlate a reference area (template) containing only a few unit cells (ideally a single one)

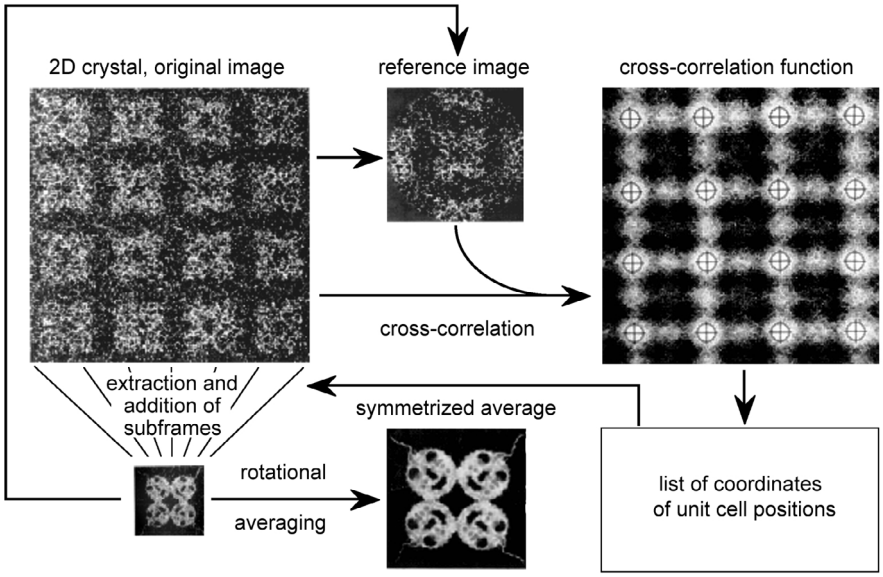


Figure 19.16 Scheme of the correlation averaging procedure for 2D crystalline objects. A reference subarea is cross-correlated with the complete image. The 2D correlation function contains bright peaks where the reference (template) was similar to the respective subarea in the original image. The peak positions are marked with crosses. These coordinates are used to extract small subareas of the image that are added (averaged) afterwards. The periodic signal is enhanced and arbitrary contributions of noise are attenuated to a (nearly) structureless average gray value. The (symmetrized) average now shows fine structure that was obliterated by noise. The average may serve as a much better defined reference for a refinement cycle.

with the entire 2D crystal (Figure 19.16). Formally, the template is centered on pixel i of the image, and we calculate the cross-correlation between the reference and the corresponding image area. The correlation coefficient r_i is written into pixel i of an array of image size (correlation function). Now the template is shifted by one pixel and the procedure repeated until the correlation coefficients of all positions are known and the correlation function is complete. The computer algorithm uses a much faster way by multiplying the Fourier transforms of the image with the complex conjugate of the transform of the reference. Back-transformation yields the complete cross-correlation function (see auto-correlation, Section 19.4.2).

In positions with best matches between the template and the image, we obtain the highest local correlation coefficients. Consequently, the correlation function shows peaks where identical or at least similar structures (unit cells) are located in the 2D crystal. These are now extracted, exactly centered to the coordinates of the correlation peak, averaged, and, if applicable, rotationally symmetrized. The correlation average is equivalent to the Fourier synthesis of an ideal 2D crystal.

The success of the approach depends on the accuracy of unit cell alignment. A correlation peak is clearly identified if it is three times higher than the standard deviation of correlation coefficients concerning noise. The height of the correlation peak r is related to the template of size D and to the signal-to-noise ratios of the template and the actual image area (Equation 19.7):

$$r \approx (S/N)_1 (S/N)_2 D \tag{19.7}$$

The equation specifies the possibilities for optimizing the accuracy of alignment. Firstly, the size of the template may be increased. This is successful as long as lattice distortions in the reference containing several unit cells do not attenuate the accuracy. Secondly, the original reference may be replaced by an average as a new template since it possesses an improved S/N. The average may now be reduced in size (down to a single unit cell) to diminish the effects of lattice distortions.

Averaging Single Particles Protein complexes that are not crystalline but distributed randomly on the grid require positioning and alignment by correlation approaches in any case. It is a prerequisite for 2D averaging that the molecules occur in a preferred orientation on the grid or that different orientations can be clearly distinguished. This is often the case with symmetric protein complexes (Figure 19.17). Once the centers of the molecules have been determined (via picking by hand or automatically by correlation approaches) and the subframes, containing one particle each, are extracted, they must be laterally and rotationally aligned by means of a reference particle or an average in refinement cycles. The search for the relative rotation of particles can be replaced by the determination of a lateral shift with particle and

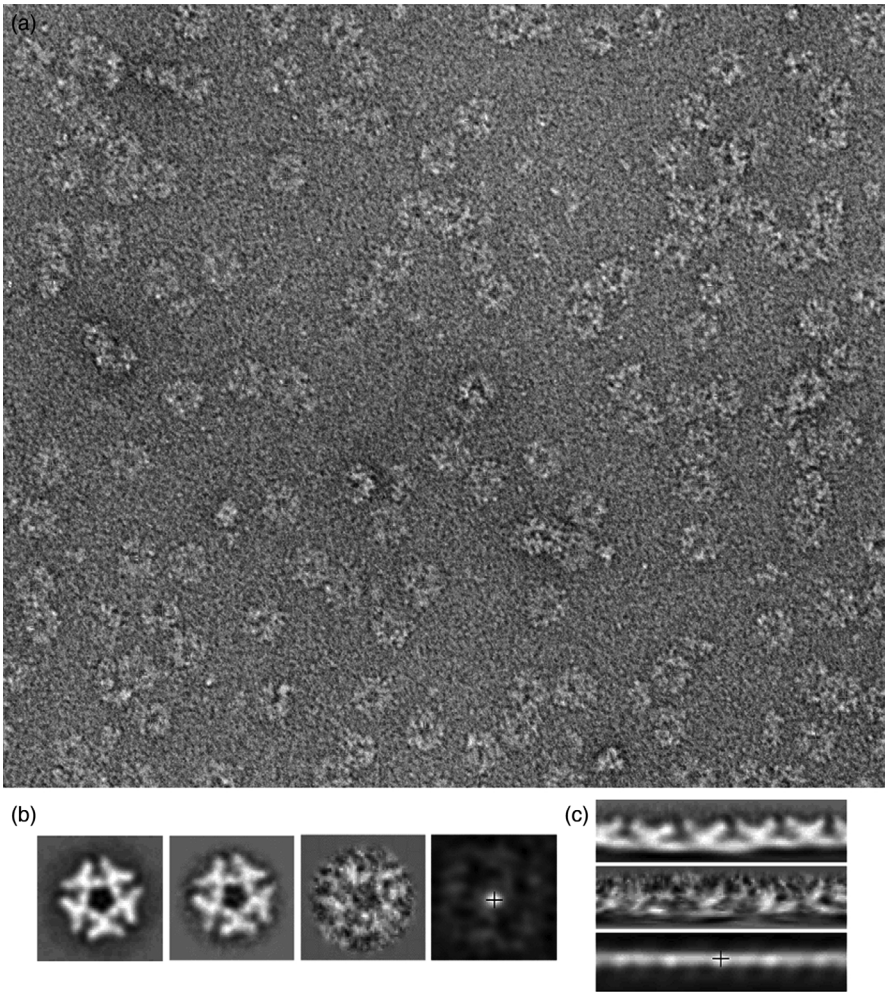


Figure 19.17 (a) Averaging of single particles with a preferred orientation on the specimen support. The manually or automatically picked particles are extracted from the electron micrograph, laterally and subsequently orientationally aligned. The latter procedure can be transformed into lateral alignment of images in polar coordinates. (b) From left to right: the five-fold symmetrized average, the non-symmetrized average (the reference for alignment), one particle extracted from the original image, and the cross-correlation function of this particle with the reference for lateral alignment. (c) The reference (average) in polar coordinates, the correspondingly transformed single particle after lateral alignment, and the cross-correlation function (bottom) indicating the peak position that corresponds to the correct rotational alignment. The protein complex, that is, the selenocysteine synthase from *Escherichia coli*, has a diameter of 18 nm.

reference data having been transformed into polar coordinates (Figure 19.17). Rotational and lateral alignments are performed successively and are repeated until further improvements make no difference.

This approach is not applicable if the protein complexes are randomly oriented, which is usually the case in cryo-preparations. These individual projections require classification into characteristic classes and (or direct) alignment in three dimensions. Averaging randomly oriented particles essentially denotes 3D reconstruction. Classification and 3D reconstruction are central procedures in image processing systems (Sections 19.4.5 and 19.5).

Principal component analysis This mathematical procedure is the basis of the well-known correspondence analysis in statistics. The latter requires data that are normally distributed and are normalized in a specific way. This criterion is not fulfilled for electron microscopical data (gray values of micrographs). They follow a Poisson distribution and are modulated by the transfer functions (see Section 19.4.3) in addition. EM data can thus hardly be analyzed by common statistical approaches and evaluated by the usual statistical significance criteria. The discrimination between random and significant image differences cannot be objectivized completely. Differences in images are considered important and meaningful if they clearly contribute to the variance and if they are structurally interpretable.

19.4.5 Principal Component Analysis and Classification

The simplest approach to distinguish different molecules or various projections of one object is to classify them by means of cross-correlation. However, this method does not tell us why the correlation is strong or weak. Distinctly different projections that both correlate badly with a reference cannot be identified by judging the correlation coefficients.

The approach of principal component analysis (PCA; single value decomposition, a variant of the statistical correspondence analysis) is a powerful method that can distinguish between different projections of objects, even those with subtle structural variations. In principle, the pixels of one image are listed in one column and the columns of all images constitute one matrix. The eigenvectors and eigenvalues of this matrix are calculated, and images containing similar structural features are thereby grouped together. However, how does the PCA find characteristic differences in a series of (up to 10^6) images? A simple example with a couple of two-pixel images illustrates the idea of the procedure.

We plot each image as a point in a coordinate system so that the gray value of the first pixel is set on the *x*-axis and the second one on the *y*-axis. If part of the images possess a darker first pixel and a brighter second one and the other part shows the opposite order the images will be distributed in a point cloud with domains that can be separated from each other more or less clearly (Figure 19.18). The mathematical procedure PCA finds the direction with the maximum variance along the cloud and orients the former *x*-axis exactly in this direction. The former *y*-axis now points in the direction of the second highest variance. The vectors, defining the orientation of the new axes, are called eigenvectors. To simplify the representation of points in the new coordinate system and to put its center into the middle of the point cloud, we subtract the average of all images from the individual ones. Hence, we are left with only those structural features that distinguish the images from each other. The new *x*-axis, called factorial axes of the first (most significant) eigenvector, no longer identifies the gray levels of only the first pixel but designates a mixed information from both pixels. The same applies to the other axis. Since the axes are still perpendicular to each other they now represent *independent structures* that are contained in the images. These are exactly those structures that result in the variances of the particular directions of the point cloud. Since the eigenvectors possess the same number of pixels as the images they can be presented as an illustration of the respective structure difference. To get a clearer idea of the structure variance described by the eigenvector, we produce two so-called eigenimages by adding and subtracting the eigenvector from the common average. The resulting images show the structure of the variation between the images (e.g., the difference of gray levels between the first and second pixel concerning the second eigenvector (factor) of our example in Figure 19.18). The eigenvalues identify the (relative) amount of variance owing to the respective factors.

The number of pixels in real EM images – and accordingly the number of eigenvectors and theoretically distinguishable structures – exceeds the dimension three by far. Multidimensional coordinate systems are not imaginable and cannot be illustrated, but they can be treated mathematically. Usually, only a limited number of eigenvectors (3–20) describe significant and interpretable structural variations. Amongst them are positional and orientational variabilities, conformational and structural differences of molecular complexes, size variations, and contrast variabilities in the case of stained samples. The majority of eigenvectors concern random variations of other sources and noise. To illustrate the distribution of images in

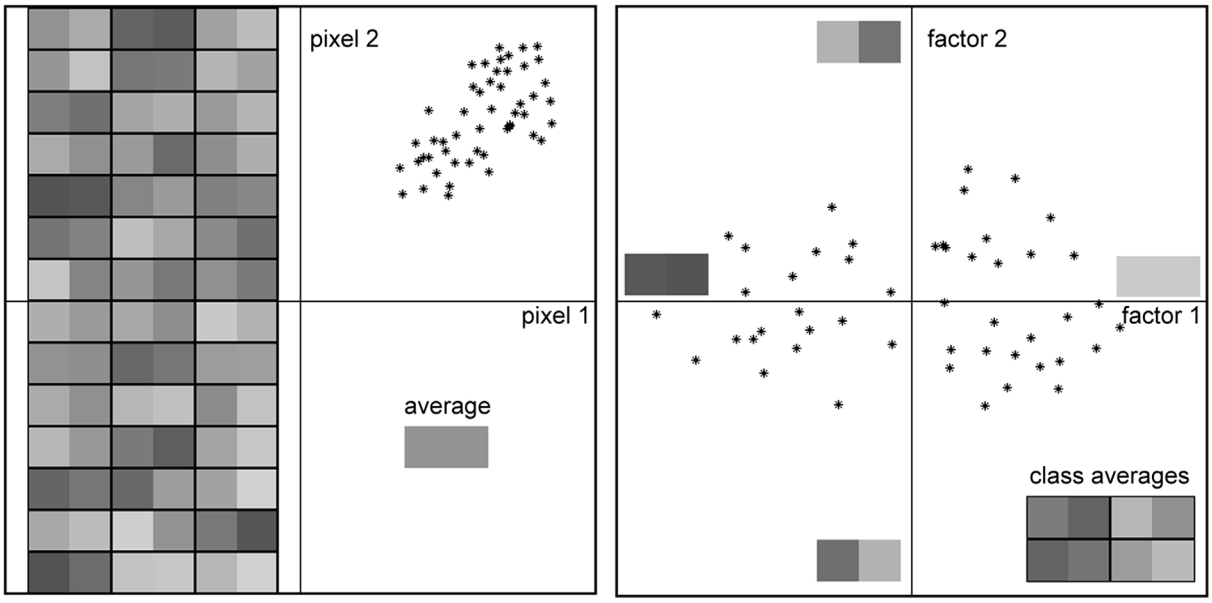


Figure 19.18 Representation of two-pixel images in coordinate systems before and after principal component analysis (PCA). (a) A gallery of synthetic images that are represented by their pixel values in a 2D coordinate system. The average of all images is structureless. (b) Distribution of images after PCA that reorients the coordinate axes according to features contributing most to the variance of images. After subtraction of the common average the origin of the new coordinate system is placed in the center of the image positions. The factorial axes now indicate structures that contribute to the pixel values; here the common gray level of the two pixels as the most significant feature (bright–dark) and the difference of gray levels between the pixels (first pixel brighter or darker than the second one). The selective averages after classification (class averages) now represent these image structures.

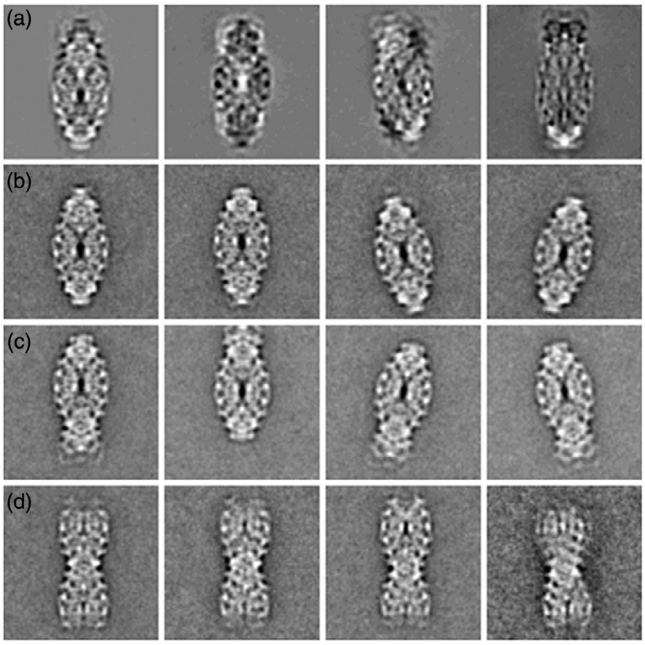


Figure 19.19 Principal component analysis (PCA) of prealigned projections of the protein complex tripeptidyl peptidase II (TPP II). (a) The first four eigenvectors significantly contribute to the variance in single particle images. The regions of highest contrast indicate pixels that are particularly strongly varying between images. These contrast variations among eigenvectors belong to structure variations that are independent of each other. The second eigenvector represents differences illustrated in the averages of TPP II in (b) and (d), the third eigenvector rotational variations with respect to the longitudinal orientation, and the fourth vector displacements along the *y*-axis and correlated orientational misalignment of 180°. (b)–(d) Class averages after PCA and classification of molecules with (b) spindle form, (c) fish-like, and (d) dumbbell projection. Further classes show longitudinal displacements, rotational misalignments of 180°, and orientational variations similar to those in the eigenvectors in (a). The eigenvectors displayed here are only a few from the whole set of projection and orientation variants.

multidimensional space we select two eigenvectors (e.g., factors 1 and 2 as shown in Figure 19.18b) and project the coordinates of all images onto this plane (i.e., we set the coordinates of the other factors to zero). This illustration is called a *factorial map*.

The definition of groups of similar images is the problem of classification algorithms that separate the point cloud in multidimensional space into distinct local domains (image classes). They use statistical procedures that minimize the sum of the squares of distances within classes, but the details are not of interest here. The class averages of our example show the expected structure differences after PCA analysis and classification, that is, dark or bright images with either a darker or brighter first pixel. Now, the structural differences are preserved much better than in the overall average, which does not show any structure. It is possible in certain cases to correct for the orientational variability of structures after PCA analysis and to eliminate the reason for some of the image variance. Applied to our example, all first pixels were now brighter (or darker) than the second ones.

Figure 19.19 illustrates the PCA analysis of real EM data that usually contain hundreds to thousands of pixels and complex structural variabilities. Here, the projections show single particles of the enzyme tripeptidyl peptidase II that were extracted from electron micrographs of frozen-hydrated preparations (Figure 19.12). The protein complexes are oriented randomly and yield a wealth of different projection images. The motivation for PCA analysis is mainly to find classes of identically oriented molecules, to correct for effects of misalignment, and to eliminate damaged complexes from 3D reconstruction. The eigenvectors (Figure 19.19a) illustrate the structural differences responsible for the distribution of images along the respective factorial axes. Regions of distinct black and white contrast indicated the differences; there, the positional and orientational variabilities are particularly prominent. The class averages (Figure 19.19b–d) illustrate the sensitivity of the PCA method in detecting misalignments that were tolerated by the correlation approach. The more precise the final alignment of projections is the better is the subsequent 3D reconstruction of the protein complex (Section 19.5.1).

19.5 Three-Dimensional Electron Microscopy

Electron microscopy of molecular complexes and electron tomography of individual biological objects both aim at a 3D reconstruction of the structures of interest. We need projections from many different directions to capture the structure as completely as possible. Randomly oriented

Table 19.1 Three-dimensional reconstruction of proteins and complex structures.

Data and methods	Single particles or individual sample			Regularly arrayed protein complexes	
	Singular complex individual sample	Many indentical molecules in arbitrary positions and orientations	Many indentical molecules in preferential orientation on the grid but arbitrarily rotated around the *z*-axis	Molecules in defined positions and orientations	Molecules in defined positions and identical orientations (2D crystals)
3D data set	Tilt series −70° to 70°	0° Projection or tomogram (intracellular molecules)	0° Projection, and one projection at ≥60°	0° Projection	Tilt series −70° to 70° with *p*1 symmetry, 0° to 70° with higher symmetries
Determination of orientation (Euler angles)	From tilt angles	Iterative fit of projections; subtomogram averaging	Rotational angles in plane by correlation, from tilt anlge	Determination from known geometric positions of molecules	From tilt angles and symmetry
Reconstruction approach	Filtered back projection or alternative approaches	Filtered back projection or alternative approaches	Filtered back projection or alternative approaches	Filtered back projection, or series expansion of orthogonal functions	Correlation averaging and Fourier synthesis
Averaging of identical structures	After 3D reconstruction, determination of positions and orientations	In the course of 3D reconstruction, or subtomogram averaging	In the course of 3D reconstruction	In the course of 3D reconstruction	In the course of averaging and 3D reconstruction
Typical specimens	Big protein complexes, amorphous viruses, ultrathin sections, and intact cells	Amorphous protein complexes (e.g., ribosomes)	Rotationally symmetric protein complexes (e.g., glutamine synthetase)	Symmetric viruses, capsids, helical structures (e.g., flagella, pili)	2D crystals (e.g., bacteriorhodopsin, S-layers, artificial 2D crystals of membrane proteins)
Resolution (cryo-preparation)	Ultrathin sections and intact cells: 1–4 nm	0.3–0.6 nm or 1–4 nm	≥0.5 nm	0.3–0.6 nm	0.3–0.6 nm

molecules in ice yield almost all orientations, whereas individual objects such as cells or cellular organelles must be systematically projected in various directions. Table 19.1 provides an overview of selected approaches for obtaining suitable orientations of specimens and methods to combine them for 3D reconstruction. The applications for single particles, regular structures, and singular objects are explained separately.

19.5.1 Three-Dimensional Reconstruction of Single Particles

Soluble protein complexes are usually randomly oriented in cryo-electron microscopy, especially when they are spherical. Consequently, one EM image already contains several different projections of the particles, and a (large) set of micrographs is likely to include all required orientations. We identify the orientation by means of the Euler angles that define how the individual coordinate systems of the molecules must be reoriented to align them to a common system. Once these angles are known we can back-project the single projections into a common volume and thereby calculate the average and reconstruct in 3D. A common approach is the filtered back projection.

The projection of a 3D object onto a 2D plane (micrograph) corresponds to a central section through the 3D Fourier transform of this object in exactly the direction that is defined by the projection angles. The Fourier coefficients are inserted in Fourier space accordingly. If, for instance, the tilt axis of tomographic projections is identical to the *y*-axis of the coordinate system the corresponding Fourier data of projections with angle Ψ_i are located in sections (planes) with angle Ψ_i with respect to the x^*-axis (to the x^*y^* plane) in Fourier space.

A projection of a sphere of diameter D will occupy a "plane" of spatial frequencies that has a thickness proportional to $1/D$ in Fourier space (Figure 19.20). The Fourier coefficients thus

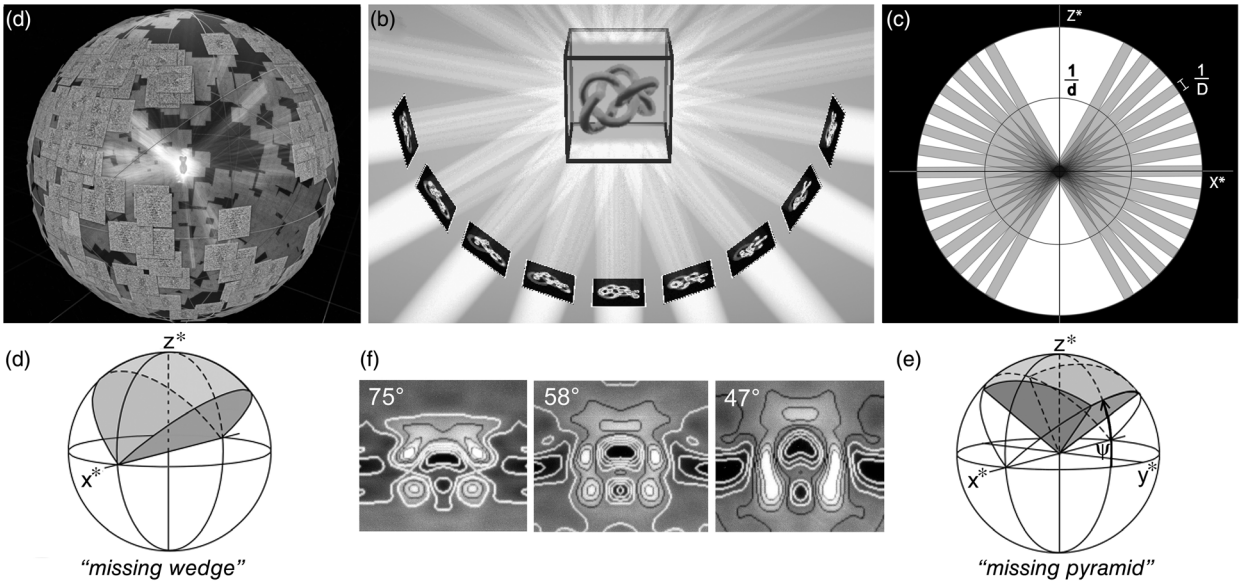


Figure 19.20 Principles of back projection and data distribution in 3D Fourier space. (a) Arrangement of projections of arbitrarily oriented single particles according to their projection angles on a (imaginary) spherical surface. Back projection of the images to the center of the sphere produces an averaged density and thereby a 3D reconstruction of the object. The reconstruction is particularly accurate if the surface of the sphere is completely covered with projections not leaving empty areas. (b) Arrangement of projections of an individual object from a tilt series and back projection into a common 3D volume in the center of the "projection circle". The latter is equivalent to a great circle (equator) on the sphere in (a). Since tilting up to ±90° is usually impossible the projection circle and the information for 3D reconstruction remain incomplete. (c) Representation of tilt series data in Fourier space (the tilt axis is identical to the y^*-axis). The projections of a spherical object with diameter D correspond to sections in 3D Fourier space with a thickness of $1/D$ that are oriented to the x^*-axis (to the x^*–y^* plane) by an angle Ψ_i, that is, the corresponding tilt angle. (d) Incomplete tilting leaves a "missing wedge" in 3D Fourier space without data (indicated in gray). (e) The space of missing data is reduced to a "missing pyramid" if a second tilt series were recorded with the object rotated by 90° around the z-axis, or if the object possesses a rotational symmetry. (f) Vertical sections through 3D reconstructions of a 2D protein crystal (see Figure 19.22) with different angular tilt ranges. The maximum tilt angle is indicated. Low tilt ranges result in low resolution in the z-direction and in an elongated object in the reconstruction.

overlap in the region of small spatial frequencies and occupy the Fourier space gaplessly to a frequency that is proportional to $1/d$. The space beyond is partially empty and would only be completely filled with data if the angular increment of tilting was reduced and the number of projections increased accordingly (Section 19.5.2). Since the coefficients below the limiting spatial frequency accumulate and artificially become larger, the data of lower frequencies are overemphasized and affect the reconstruction. To account for this effect the Fourier coefficients are weighted (filtered) in a frequency-dependent manner in the course of filtered back projection.

The electron microscopist is, however, confronted with another fundamental problem. The individual orientations of particles are unknown and so are the projection angles that are required to insert the data into the right place in Fourier space. One solution to the problem is an iterative approach that starts from a low-resolution 3D model of simple shape (e.g., a globular or cylindrical density of appropriate size). The real projections are compared with a series of calculated projections from the model. The best match yields an estimate for the projection angles, the real projection is inserted into Fourier space accordingly and replaces the respective data of the model projection. In this way the 3D data become more detailed and successively improves until, after several rounds of iterative fine-tuning, the projection angles remain constant and the 3D reconstruction reaches a final state. It may be advantageous to sort the noisy images into projection classes beforehand (Figure 19.19) and to remove damaged particles and different conformational states by PCA. There are several procedural variants of this approach but all rely on iterative optimization of the 3D reconstruction. Two factors are always helpful: the more projections are available, the higher the S/N ratio of the reconstruction becomes, and the higher the number of different projections that can be inserted the more completely is the Fourier space filled with data and the better will the 3D model be defined.

The final step of the reconstruction procedure is the inverse Fourier transformation of the 3D data into a real image cube. The result is a x-y-z volume of 3D image elements (voxels) that represent the density values of the molecular structure. Single x-y- (x-z- or y-z-) or arbitrarily

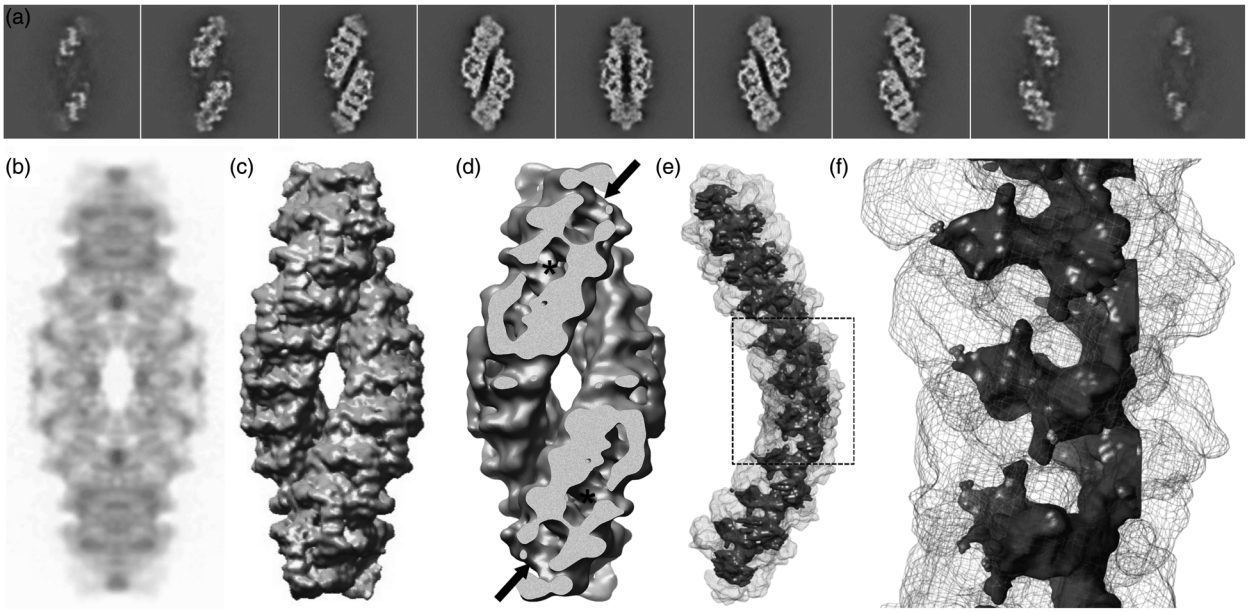


Figure 19.21 Three-dimensional reconstruction of the enzyme complex tripeptidyl peptidase II from *Drosophila melanogaster* by single particle averaging. (a) Selected slices (*x*–*y*) of the 3D data cube showing the enzyme in longitudinal orientation. Size of images 67 nm × 67 nm; distance between slices 2.6 nm each. Bright regions represent protein, dark areas in-between cavities in the enzyme. (b) Projection through the 3D data volume, (c) surface rendered 3D model, (d) 3D model cut open to illustrate the molecular cavities, and (e) 3D model of the total cavity system (black) of one half of the protein complex built of ten dimers. (f) Enlarged model of the cavity system that contains the active centers of the protein subunits. The enzyme is 60 nm long and has a maximum diameter of 30 nm. Source: courtesy of Beate Rockel, Martinsried, Germany.

oriented slices may be extracted for analysis and illustration. These density data provide the most objective presentation of reconstructed structures. A surface-rendered model (Figure 19.21) provides a very good illustration of the 3D structure of the volume; it is thus a popular model. The surface is defined by a threshold of the density (gray) values that identifies the border of the object (protein) in distinction to the surrounding medium (ice). These models easily mislead us into to assuming that they represent the ultimate 3D structure. However, due to principal restrictions, it is difficult to identify the correct threshold exactly. One restriction is that the resolution limits the accuracy of smaller details and prevents the definition of a sharp molecular border in the reconstruction. The mass–volume relationship of proteins (Equation 19.4) may help to build reasonable models.

Sometimes, macromolecular complexes with internal symmetry assume a preferred orientation on grids or even in frozen-hydrated preparations. Typically, flat complexes lie in a parallel plane to the carbon support or the surface of the ice film and thus limit the variability of orientations (Figure 19.17). However, the complexes are free to rotate around an axis perpendicular to the grid (*z*-axis) and thus assume random orientations. In these cases, two micrographs are sufficient for 3D reconstruction (Table 19.1). The 0° projections serve to determine the rotation around the *z*-axis in the plane (Section 19.4.4), and a projection of the tilted grid yields the complete 3D data set since the randomly rotated molecules produce different projections. The maximum tilting angle is limited to 60–70° (90° would be ideal), and a cone-like volume without data around the z^*-axis in Fourier space remains, that is, the *missing cone* (Figure 19.20). Again, the consequence is a reduced *z*-resolution of details with low spatial frequencies in *x*-*y*. It is sometimes helpful to vary the ionic milieu of the sample to induce another preferred orientation on the grid and to complement for missing data.

19.5.2 Three-Dimensional Reconstruction of Regularly Arrayed Macromolecular Complexes

Protein complexes in a regular 2D lattice are identically oriented and are all viewed in the same way in the microscope. We collect different projections by tilting the object around an axis and

recording a series of N images at defined tilt angles. But how many projections or which angle increment $\Delta\Psi$ do we need to resolve structural details of size d? Looking at the data in Fourier space (Figure 19.20) and assuming that they are distributed over the whole angular range, R.A. Crowther derived Equation 19.8. Since tilting to 90° is usually impossible, the remaining *missing wedge* impairs resolution in the z-direction (Figure 19.20):

$$N = \frac{\pi D}{d} \tag{19.8}$$

$$\Delta\psi_i = \frac{d\cos\psi_i}{D} \tag{19.9}$$

It is advantageous for 2D crystals to adjust the tilt angles not in an equidistant manner but to successively decrease the angular increment with higher tilt angles. Now, the data are equidistantly distributed in z^*-direction in Fourier space, which is of particular value for high-resolution structures. Two-dimensional crystalline monolayers produce reflexes in x^*-y^* in Fourier space and continuous lines (lattice lines) in z^* in these places since there are no repetitive structures in the third dimension. Projections with constant angular increments "cut" these "lines" with increasing distances in z^* and yield data that are separated by increasing gaps. Interpolation in the course of reconstruction produces equidistant data. These interpolations are particularly accurate if the distances between measured data are small and constant. The required angular increment following tilt angle Ψ_i (rad) beginning at 0 is described by a function of the desired resolution d and the thickness D of the crystal. It is given by W.O. Saxton's equation (19.9). Since 2D crystals usually show rotational symmetry (except for space group $p1$) it is sufficient to record series with positive or negative tilt angles only (0° to >60° or to <−60°). Rotational symmetry also reduces the region of missing data from a missing wedge to a missing pyramid (Figure 19.20). Two-dimensional crystals of macromolecules are usually not ideally regular. It is thus profitable to average the unit cells by correlation methods and to then combine the averages in Fourier space. Averaging and interpolation of lattice lines efficiently remove superimposed noise.

Biological objects with several or many identical macromolecular units in regular arrangements – for example, capsids of viruses and phages or in helical structure such as flagella, pili, and other protein filaments (Table 19.1) – are suited for averaging approaches. If the geometrical positions of repetitive units are known, for example, the helical arrangement and the number of units per helical turn, it is possible to define the projection geometry for all units. The great advantage of these macromolecular structures is that, in principle, a single projection is sufficient to calculate the 3D reconstruction since one micrograph already contains various projections of the macromolecule. The same goes for virus capsids. In addition, it is of course possible to combine a couple of 3D data sets for a final, better-defined reconstruction (Figure 19.22).

Tomography All tomographic reconstruction approaches share the same principles, that is, projection of the object density by transmission imaging, combining projections from different directions, and calculation of a three-dimensional data cube. The methods are non-invasive and provide insight into an intact object. Essentially, all electron microscopical 3D reconstruction approaches are tomographic ones. However, it has been vernacularized to speak of tomographic reconstructions of individual, non-crystalline objects (complex protein assemblies or intact cells) that were projected under different angles in a tilt series.

19.5.3 Electron Tomography of Individual Objects

Individual biological objects such as macromolecular assemblies, membranes, amorphous viruses, ultrathin sections of cells, cell organelles, or intact cells require tilt series data for 3D reconstruction. This application is known as electron tomography, with reference to computer tomography of macroscopic specimens or probands.

Three-dimensional reconstructions of single molecules and regular structures, such as described in the previous sections, always include averaging steps of identical particles to increase the quality and completeness of data for the final 3D model. But the reconstruction process itself – that is, filling the Fourier space with data and back-projection or equivalent approaches – is independent of averaging. It is also possible to reconstruct non-repetitive, individual objects (Figure 19.20b). Indeed, the specific feature of electron microscopy is its ability to image individual structures down to the subnanometer range. To do so, (cellular) cryo-electron tomography has to solve two incompatible challenges. On the one hand, we should record many projections over the whole angular range with maximum signal-to-noise ratio to achieve the best resolution of small structures. On the other hand, it is necessary to minimize the

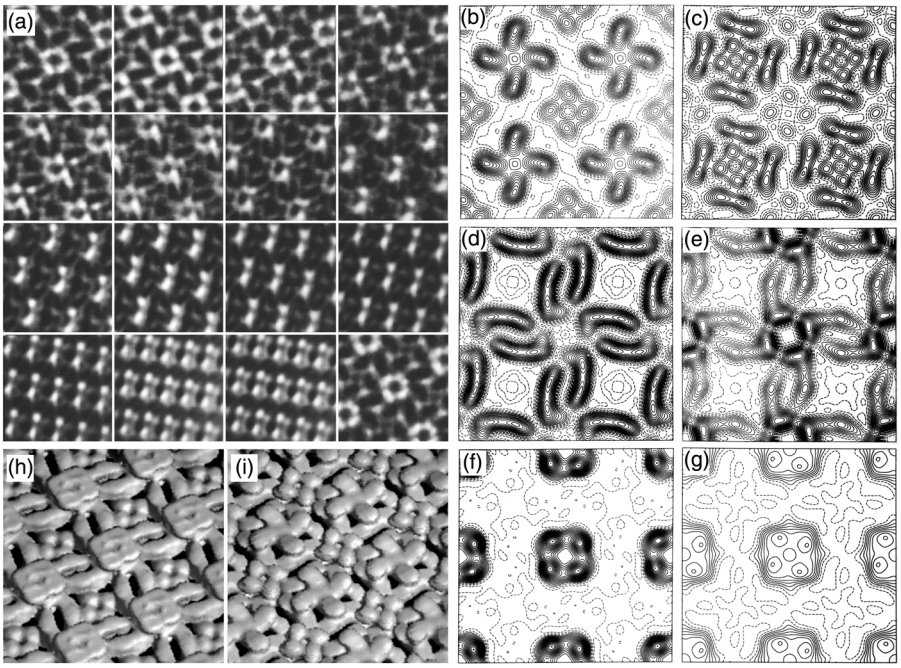


Figure 19.22 Three-dimensional reconstruction of a negatively stained 2D protein crystal, the surface layer (S-layer) from the Gram-positive bacterium *Sporosarcina urea*. (a) Averages of the tilt series projections with non-equidistant tilt angles from 0° to 78°. (b)–(g) Horizontal sections through the 3D reconstruction in contour line presentation. The 3D volume contains four unit cells with the lattice constants $a = b = 13$ nm and the lattice angle of 90°. The sections show the protein density from the outer surface of the layer (b) to the inner surface (g) in positions at −2.6, −1.2, −0.3, 0.3, 1.2, and 2.5 nm with respect to the central plane. (h) and (i) Surface-rendered 3D models showing the inner and the outer surface of the protein layer. For vertical sections through the central protein domain see Figure 19.20.

total electron dose so as not to just destroy the structures that we want to see in the reconstruction!

Fractionation of the total dose over all micrographs of the tilt series results in a corresponding loss of the S/N ratio with the consequence that structures may not be detectable among the obliterating noise. The S/N ratio of individual projections must be so high that they are aligned with each other (Section 19.3.3). The mechanical accuracy of the goniometer is not good enough to avoid image shifts upon tilting. These unavoidable errors have to be corrected for afterwards. Small gold clusters (colloidal cold) that are added to the sample before freezing and provide high contrast anchor points facilitate the correlation alignment. Now, the S/N of the structures of interest is not limiting for alignment and may even be low. The two conditions – maximally allowed cumulative electron dose and minimal contrast (especially in images taken without a phase plate) – limit the possible number of projections of a tilt series and thus the achievable resolution of an object of thickness D. Cellular objects should not be thicker than ≈0.5 µm, which means that cryo-electron tomography applications are restricted to viruses, a number of prokaryotes, and flat areas of eukaryotic cells in the first place. However, cryo-microtomy and ablation of material by ion beam treatment (FIB) offer ways to investigate sections or subvolumes of bigger biological cells in a close-to-native state (Section 19.2.1).

Reconstructions of individual objects miss the opportunity to minimize noise by averaging. Other criteria are required to separate structures from uncorrelated noise. One approach is *nonlinear anisotropic diffusion*. This algorithm exploits the moderate variation of voxels belonging to a continuous structure, which is different from the randomly varying and uncorrelated values of noise.

The missing wedge of data also creates problems in tomograms (Section 19.5.2). One effect is that flat (thin) structures such as membranes become blurred or even invisible in the z-direction. Accordingly, the reconstructed top and bottom regions of cells are incomplete. To reduce the region of missing data in Fourier space one could rotate the grid by 90° in the microscope after recording a tilt series and take another series. The missing wedge then becomes a missing pyramid and the data is more complete (Figure 19.20).

The final result of a 3D reconstruction is a data cube that contains the density distribution of the reconstructed cell or section. The interpretation of these complex and detail-rich structures requires special approaches for analysis and visualization that allow the researcher to identify and visually enhance distinct structures (Figure 19.23 and Section 19.6).

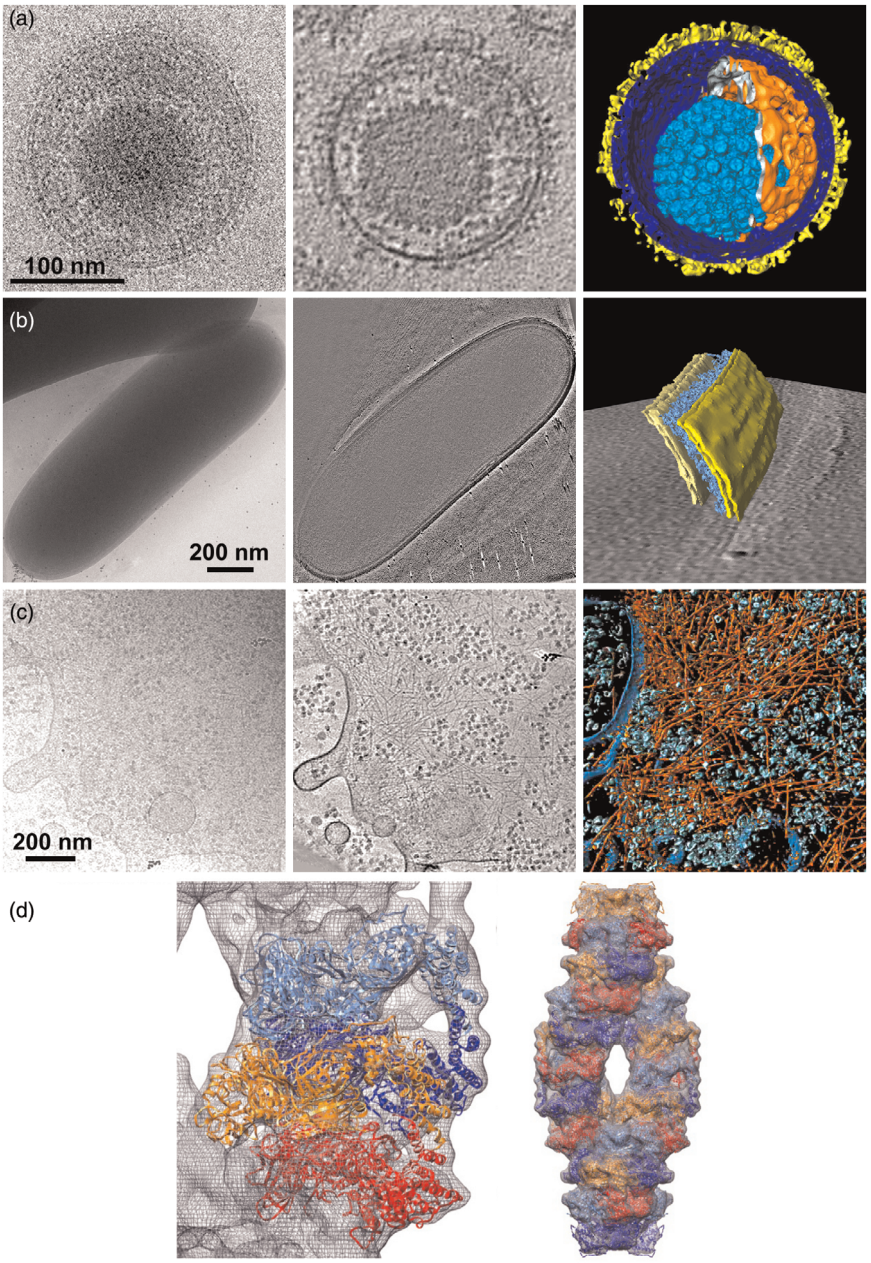


Figure 19.23 (a) Original projection and reconstructed 3D data of the virus *Herpes simplex*, (b) of *Mycobacterium bovis* BCG, and (c) of a part of the eukaryotic cell *Dictyostelium discoideum*. All the organisms were imaged by cryo-electron tomography without any chemical fixation and staining and reconstructed by the approach of filtered back projection. The images show the 0° projection, one (central) *x*–*y* slice of the reconstruction and the surface-rendered 3D model. The model of the mycobacterium shows the lipid bilayer structure of the inner and outer membranes and material in the periplasmatic space, that is, cell wall polymers. (d) Surface model of the tripeptidyl peptidase II from *Drosophila melanogaster* as obtained from single-particle cryo-electron microscopy; the protein dimers from X-ray structure determination are fitted into the electron density. This hybrid approach allows us to calculate a quasi-atomic model of the entire 40-meric enzyme complex. Source: Courtesy of Kay Grünewald (a), Ohad Medalia (c), and Beate Rockel (d), Martinsried, Germany.

19.6 Analysis of Complex 3D Data Sets

19.6.1 Hybrid Approach: Combination of EM and X-Ray Data

Cells contain large and heterogeneous protein complexes that may be purified without losing their integrity but that do not crystallize for X-ray structure determination. Examples are the 26S proteasome, the spliceosome, polysomes, the bacterial flagellar motor, the nuclear pore complex, centriole structures, cilia, the cytoskeleton network, and many others. However, in many cases it is possible to isolate the subunits and to determine their atomic structures individually. These subunits can then be fitted into the density of the intact complex as obtained

from single particle or tomographic reconstructions and can be used to calculate a pseudo-atomic model. This hybrid approach often is the only successful way to get deeper insight into the structure and functional conformations of protein complexes; the enzyme TPP II is a typical example (Figures 19.21 and 19.23).

There are two approaches for fitting, namely, *rigid body docking* and *flexible fitting*. Rigid body docking is suitable for moderate resolution structures (1–3 nm) where the best place and orientation of subunits is determined by correlation methods. Higher resolution structures often reveal local discrepancies between the atomic structure and the reconstructed model that derive from real conformational differences. Molecular dynamics calculations allow adaptation of the atomic structure in such a way that it fits into the structure of the whole complex.

Meanwhile, there are explorations of approaches to model structures consisting of many parts or structurally unknown components, such as the nuclear pore complex. Here, the combination of biochemical data and information from other sources yields additional criteria to define restricting conditions for modeling.

19.6.2 Segmenting Tomograms and Visualization

Tomograms of cells or cellular segments contain the signatures of many (theoretically all) macromolecules and cellular structures one would like to analyze and identify in 3D presentations. Due to the high protein concentration in living cells (80–400 mg ml^{-1}), creating the phenomenon of *macromolecular crowding*, it is difficult or even impossible to visually assign density data in tomograms to distinct structures. It is thus necessary to demarcate structures of interest such as membranes, filaments, or macromolecular protein complexes from each other. They are *segmented* and usually highlighted by colors in 3D models. Figure 19.23 shows EM data and the corresponding 3D models of segmented tomograms from a virus, a bacterial cell, and a segment of a eukaryotic cell.

The simplest means of segmentation is to define a gray value that represents the border of a molecular surface and to blank out all voxels below this threshold. This is the common method of rendering 3D models of single molecules (Section 19.6.1), but it is not suitable for complex structural assemblies such as biological cells. Here, we need criteria to identify voxels that are correlated and belong to a coherent structure. There are automated procedures that recognize membranes and filaments even in noisy 3D data, but it will sometimes be necessary to refine segmentations by hand. Segmentation of single protein complexes in the crowded cytoplasm is not possible and so we need more powerful approaches.

Macromolecular crowding The high concentration of macromolecules in cells means that about 30% of the cytoplasmic volume is occupied by protein complexes and other big biological molecules. The characteristic distance between neighboring macromolecules is only 10–20 nm, that is, the size of many protein complexes. Big complexes can thus not occupy these intermolecular zones – only a rather limited volume remains available for them. There is literally a shortage of space in crowded cells. This situation causes dramatic equilibrium shifts for reactions that increase or decrease the occupied volume (including the surrounding space) of macromolecules. Amongst these processes are isomerization and folding reactions or oligomerization and dissociation of protein complexes. This is the reason why supramolecular complexes and protein assemblies are more stable in cells than in the diluted environment *in vitro*. Some of these (hypothetical) complexes can thus only be observed in intact cells.

19.6.3 Identifying Protein Complexes in Cellular Tomograms

If protein complexes are located closely together in cells or if they are part of a supramolecular structure we use the approach of template matching to identify and localize them in tomograms (Figure 19.24). For this purpose, we need a 3D model of the protein complex of interest. Models from X-ray crystallography or single particle reconstructions are suitable data. Cross-correlation of the tomogram with the template yields the position and the orientation of the molecules of interest. The correlation is a function of the two 3D data sets and the six degrees of freedom for the position (x,y,z-coordinates) and the three Euler angles. The required computing power is challenging. However, the approach is feasible; experiments have localized and identified different molecules in phantom cells (lipid vesicles) (Figure 19.24). These experiments showed that a resolution of at least 2 nm is necessary to identify the target molecules reliably and to minimize false positive hits. This is a challenging proposition, but the new electron detectors and improved imaging conditions promise fruitful scenarios. The 3D distribution and arrangement of ribosomes in bacterial cells have already been investigated (Figure 19.24), the native structure of polysomes (poly-ribosomes) and inactive pairs of hibernating ribosomes have been studied *in situ*, and recent research has identified active and inactive conformations of the 26S proteasome in neuronal cells.

Tomograms of cells are actually unique structures, but they contain several redundant proteins that may be extracted from the 3D data set after localization and identification and that

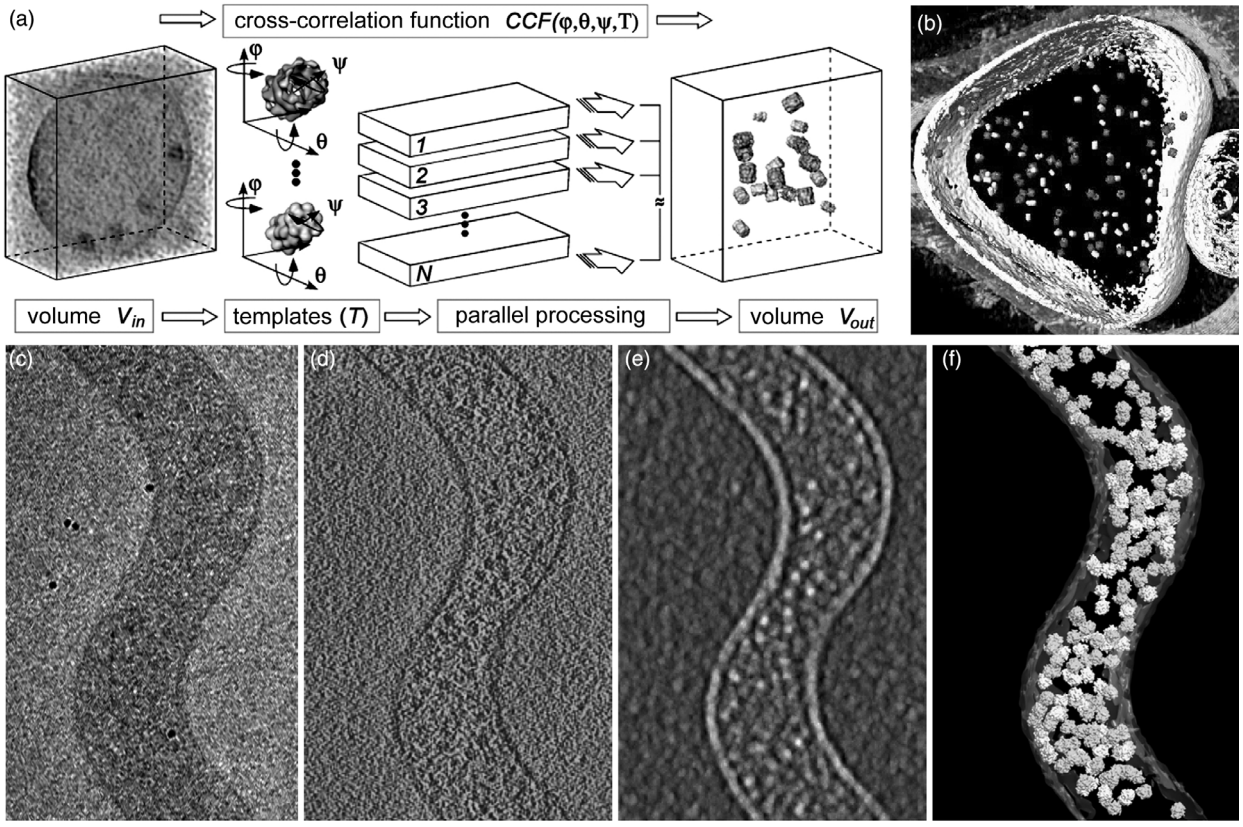


Figure 19.24 (a) Detection of macromolecule complexes in a 3D data set (tomogram) of a cell is performed by cross-correlating the reconstructed data with a 3D template (template matching). The correlation determines the spatial position (*x, y, z*) and orientation (three Euler angles) of the molecules. The elaborate process is calculated in a parallel computer. (b) Positions and orientations of the protein complexes proteasome (bright) and thermosome (gray) in a "phantom cell" (lipid vesicle). (c) Projection of the vitrified bacterium *Spiroplasma melliferum*, (d) slice of the tomogram that shows large protein complexes in the cytoplasm, and (e) image of the correlation function obtained from the tomogram and the template 70S ribosome; bright spots reflect the identified position and orientation of ribosomes in the 3D data set of the cell. (f) 3D model of the cell containing models of the ribosome at places and in orientations as determined by template matching. Size of image 600 nm. Source: parts (c)–(f) courtesy of Julio Ortiz, Martinsried, Germany.

may be averaged after classification (*subtomogram averaging*). The appeal and the scientific potential of this process is that the complexes remain in the natural environment, in a native and untouched state, and that their functional interactions with surrounding proteins can be visualized. The nuclear pore complexes are one example: they are immobilized by freezing in different situations of translocating the cargo. By cumulating many subtomograms and sorting them into the right order it was possible to obtain a "movie" of the translocation process.

19.7 Perspectives of Electron Microscopy

About 70 years after its invention in 1931 electron microscopy opened up new perspectives for molecular structural biology and cytology by establishing advanced cryotechniques, 3D reconstruction, and visualization methods. About a decade later, two technical (r)evolutions mark another "quantum step" in electron microscopy – new electron detectors, improved resolution, and an employable phase plate ameliorated contrast. Two lines of applications delineate the fascinating future of electron microscopic structure research.

Single particle analysis of protein complexes deals with many projections, reaching up to millions. With the use of direct electron detectors it is now realistic to aim for the atomic resolution of complexes, especially of those that cannot be tackled by other methods of structure research. While it is possible to obtain atomic models of rigid protein complexes with less than

10^5 single particles, an even more fascinating perspective is to collect as many projections of flexible complexes in various conformational states as possible, to then classify them, and to sort the different structures into a consecutive order of transformation. The result is a quasi-time- or process-resolved series obtained from "four-dimensional" electron microscopy. The first examples were the translocation of tRNA between different binding sites in ribosomes and the visualization of the flexibility of the nuclear pore complex in native nuclei. Investigations at higher resolution showed conformational changes of the 26S proteasome that led to an atomic model of functional transitions.

The development of *cryo-electron tomography* paved the way for 3D models of intact cells or sections thereof in a close-to-life state. Even at a moderate resolution of 3–5 nm we can detect intracellular structures that cannot be observed in conventionally fixed and embedded preparations. This immediately shows the significance of CET for the investigation of pathological effects in cells. The improved detectors, the correction of beam-induced movements, and the impressively enhanced image contrast created by a phase plate mark a further quality step. We can realistically expect to interpret cellular tomograms at the level of about 1 nm in the near future. A single tomogram contains the signatures of hundreds of proteins and supramolecular complexes, that is, a wealth of 3D information of the cellular proteome. We assume that many proteins interact in the cytoplasm and temporarily form supramolecular complexes under the conditions of macromolecular crowding in cells. Such complexes tend to dissociate in diluted environments and cannot be isolated as stable structures. Only cryo-electron tomography enables us to visualize macromolecular aggregates *in situ* and to investigate the structural network of different macromolecules. Once the proteins in tomograms are identified and their positions and orientations known, it is possible to dock atomic models into the electron densities and to create a pseudo-atomic map of macromolecular structures and their interactions in individual cells. There is still some way to go until medium-sized protein complexes or even small ones can be unambiguously identified, but we have taken the first steps.

Further Reading

Frank, J. (2006) *Electron Tomography: Methods for Three-Dimensional Visualization of Structures in the Cell*, 2nd edn, Springer, Berlin.

Frank, J. (2006) *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in their Native State*. Oxford University Press, Oxford.

Hawkes, P.W. and Spence, J.C.H. (2007) *Science of Microscopy*, vol. I, Springer, New York.

Reimer, L. and Kohl, H. (2008) *Transmission Electron Microscopy*. Springer Series in Optical Sciences, vol. 36, Springer, New York.

Williams, D.B. and Carter, C.B. (2009) *Transmission Electron Microscopy. A Textbook for Materials Science*, 2nd edn, parts 1 to 4, Springer, New York.

A public domain program for the analysis and processing of microscopic images: NIH, ImageJ: Image Processing and Analysis in Java. http://imagej.nih.gov/ij/

Atomic Force Microscopy

20

Daniel J. Müller[1] and K. Tanuj Sapra[2]

[1]*ETH Zürich, Biosystems Science and Engineering, Mattenstrasse 26, 4058 Basel, Switzerland*
[2]*Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland*

20.1 Introduction

The year 1986 ushered in a new era of imaging and manipulation of hard and soft matter at the nanoscale, turning theoretical (or undreamed of) possibilities into practical realities. The invention of a conceptually simple yet very powerful device, the atomic force microscope (AFM), provided a tool to zoom into the molecular scale thereby revolutionizing nanotechnology. The last three decades have witnessed great strides in AFM technology; it is now a routine to map objects with a resolution of up to a few angstroms (Å), manipulate them with high-precision, and at the same time quantify their physical, chemical, and biological properties.

The AFM belongs to the family of scanning probe microscopes (SPMs), which utilize a sharp probe as a scanning "stick" and a handle to manipulate (e.g., pick, drop, remove) objects at the nanoscale. The SPM microscopy technique relies on specific interactions between the probe and the object; in many cases the interactions can be tailored to a specific sample or the application. The detection system in SPM microscopes includes optical signals in scanning near field microscopy (SNOM), tunnel currents in the scanning tunneling microscope (STM), ion currents (scanning ion conductance microscope, SICM), or magnetic interactions in the magnetic force microscope (MFM). The detection versatility has enabled the development of more than 20 different measurement applications for the scanning probe microscopy of inorganic and organic samples.

Electron Microscopy, Chapter 19

As explained in Section 20.2, the AFM detects interaction forces between a sharp atomic or molecular probe and the object. A major advantage of the AFM is that sensitive biological samples can be investigated in their natural aqueous milieu under defined conditions; for example, specific pH, ion compositions, and temperature, which simulate the physiological conditions. The signal-to-noise ratio of the AFM compares superlatively to the optical and electron microscopes. However, a thorough understanding of the molecular interactions between the AFM cantilever tip and the sample (Section 20.3) and optimum sample preparation (Section 20.4) are imperative prerequisites for achieving a signal with low noise to enable molecular mapping of soft biological matter at sub-nanometer resolution. Because the energy imparted by a cantilever tip while scanning a surface (Section 20.3) is of the order of thermal energy ($\approx 3.5 k_B T$), individual biomacromolecules can be scrutinized with high precision *in vitro* or *in situ* without compromising their structural and functional integrity. In contrast, photons of wavelength ≈300 nm possess energies $\approx 3150 k_B T$, which is sufficient to break covalent bonds of organic molecules and is detrimental to protein structure.

Section 20.5 provides a few examples of sub-structural (high-resolution) AFM imaging of biological cells, single proteins, nucleic acid polymers, and sugar chains under native conditions. The fast recording of imaging sequences allows direct observation of the molecular machineries at work, and the dynamics of the macromolecular complexes in a cell. Imaging

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

with AFM extends beyond simple topographic mapping; information on the physical and chemical properties of a system can also be obtained by functionalizing the cantilever probe. Of direct relevance to studying biological mechanisms is the monitoring of various biological interactions, for example, cell–cell adhesion, interaction forces between individual receptor–ligand complexes, chemical groups, or even observing the convolution and unfolding processes of individual proteins (Section 20.6). In addition, the AFM is gaining importance as an analytical tool; for example, to determine the mechanical properties (flexibility, stability) of biological and synthetic polymers, and also biological macromolecules. It is becoming increasingly evident that the AFM offers unprecedented high spatial resolution unmatched by any bioanalytical or biophysical approach. Thankfully, the potential of scanning probe microscopy in bioanalytical applications is far from being exhausted.

20.2 Principle of the Atomic Force Microscope

Inarguably, the most crucial functional element of the AFM is a sharp probe (or tip) made of silicon or silicon nitride (Si_3N_4) and engineered at the ventral end of a microscopic spring lever (also known as the cantilever); the whole system is designed to have a low spring constant ($k \approx 0.1\,N\,m^{-1}$) (cantilevers with higher spring constants are produced for applications in material sciences). For the cantilever to function as a force sensor, the dorsal or the top surface of the cantilever is usually metal-coated onto facilitate and maximize reflection of the laser beam to a quadrant photodiode (Figure 20.1). The quadrants of the photodetector maintain a voltage difference, which to a first approximation depends linearly on the applied force. The inclusion of a piezoelectric transducer optimizes the raster scanning capabilities of the tip, which is able to inspect each defined point of the biological surface while detecting the interaction forces at those points (Figure 20.1). The cantilever acts as a force amplifying arm, that is, upon sensing a force the cantilever deflects and the signal is registered on the location-sensitive photodetector. The interaction forces consist of repulsive and attractive contributions, and are typically in the range 0.01–100 nN (Table 20.1).

Table 20.1 Forces between the AFM tip and object.

Force	Direction of force	Range
Pauli repulsion	Repulsive	Extremely short (≤0.2 nm)
Van der Waals interactions	Attractive	Very short (a few nm)
Electrostatic interactions	Attractive/repulsive	Short (nm to μm)
Capillary (probe in water)	Attractive	Long (μm to mm)

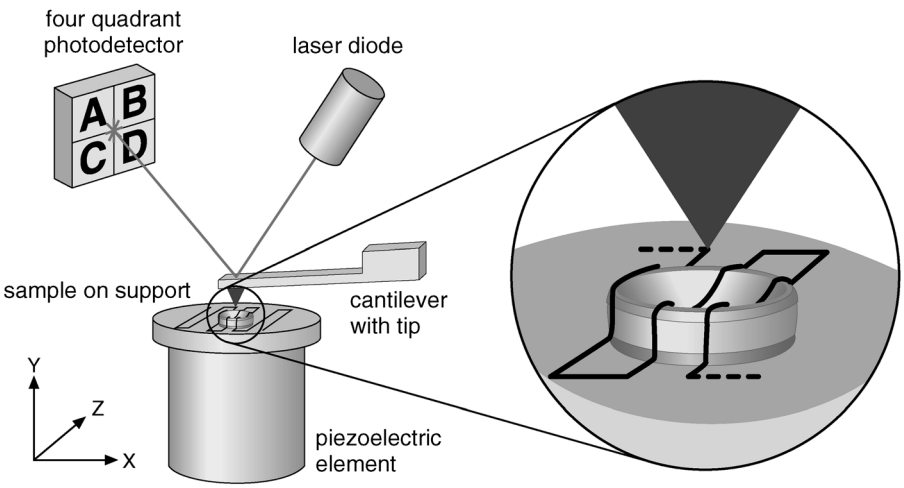


Figure 20.1 Schematic representation of the atomic force microscope. An immobilized sample is moved by means of a triaxial (*x*, *y*, *z*) piezoelectric element under a sharp scanning cantilever probe. During this raster movement, the bending of the cantilever spring is measured by a laser beam reflected onto a photodetector. The voltage difference between the upper and lower segments of the photodetector ($V=(A+B)-(C+D)$) is a direct measure of the cantilever spring deflection, which is used to calibrate the spring constant and quantify the minisscule interaction forces.

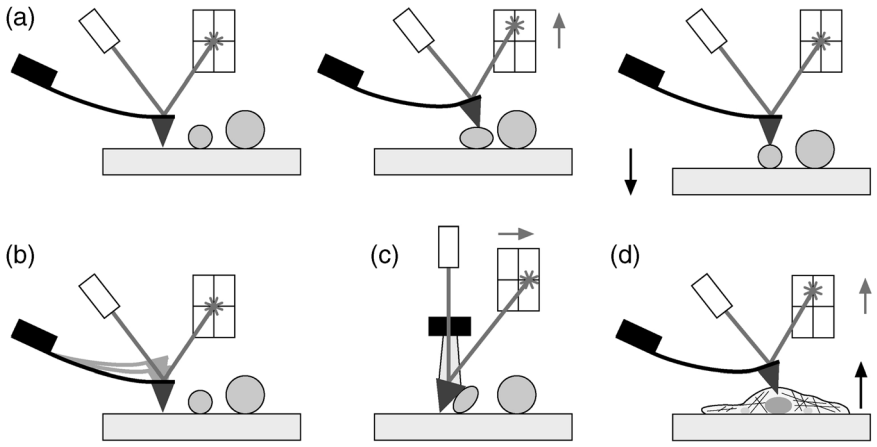


Figure 20.2 Imaging with atomic force microscopy. (a) In the contact mode, the tip is in continuous contact with the biological sample and, therefore, follows the surface features of the sample even upon encountering an obstacle, resulting in deflection of the cantilever spring. To protect the sample from excess force, the contact force (measured by cantilever deflection) is maintained by changing the Z-position of the sample with the aid of a feedback loop. The error signal is used to map the sample topography. (b) In the dynamic mode, a cantilever is oscillated close to its natural resonance frequency with an amplitude of a few nanometers. Thus, the object is touched only briefly during the downward travel of the cantilever avoiding lateral interactions. The feedback control loop in this case is used to maintain the oscillation amplitude. (c) Lateral scanning of the AFM tip can deform a soft biological object. This can be prevented by precise adjustment of the contact force, the scan speed, and the feedback parameters. (d) Elasticity is determined by moving the cantilever tip against a soft object (e.g., cells). The ratio of the distance travelled by the cantilever into the sample and the cantilever deflection associated with this gives a measure of the sample elasticity.

So far, several methods have been developed for scanning force microscopy. In the most commonly used scanning method – the contact mode – the cantilever tip is maintained in contact with the sample surface at a constant user-defined force. This is achieved by keeping the cantilever deflection (i.e., the force) constant by continuously monitoring and changing the distance of the sample surface from the cantilever tip. To ensure the integrity of the biological samples and that the surface structures are not irreversibly deformed, a scanning force of $\leq$0.1 nN is generally used (see Section 20.3). By regulating the height required for a constant contact force, the surface features of the sample are mapped point-by-point resulting in a surface topography (Figure 20.2a).

Besides forces normal to the plane of the sample surface, lateral forces can deform or scratch away a soft biological object from its base during raster scanning in contact mode (Figure 20.2c). The undesirable sample distortions by normal and lateral interactions can be minimized by setting low contact forces (see Section 20.3). Dynamic imaging methods (e.g., Tapping™ mode) offer an alternative approach to reducing the impact of the scanning tip on the object. In the Tapping™ mode, the cantilever spring is stimulated to a sine-wave oscillation close to its natural resonance frequency. The oscillation is a means to ascertain that the tip touches the biological object only at the lower end of each cycle (Figure 20.2b), as a result minimizing the contact time and lateral interactions. Dynamic imaging procedures are therefore particularly suited for imaging weakly immobilized biological objects (e.g., cells or fibrillar structures). A crucial criterion for obtaining an unperturbed topography is to maintain the oscillation amplitude of the cantilever by controlling the distance between the cantilever tip and the sample surface.

Topography Three-dimensional illustration of a surface.

20.3 Interaction between Tip and Sample

Akin to its siblings of the SPM family, the AFM can deliver resolution up to atomic dimensions. Albeit the achievable resolution depends crucially on the tip sharpness and the surface corrugations of the object, the finite radius of the cantilever tip limits the contouring of sharp and delicate features of an object. As shown in Figure 20.3a, the tip can broaden the lateral dimensions of an object. In contrast, structural periodicities with less defined edges can be reproduced with a higher accuracy (Figure 20.3b). In all these cases, the measured topography represents a nonlinear superposition of the object under investigation, the strength of the pronounced details being dependent on the corrugation of the object and the tip dimensions.

Owing to the considerably soft nature of biological objects compared to the cantilever spring, contact forces of $\geq$1 nN exerted by a cantilever tip are sufficient to deform and denature proteins. As explained above (Section 20.2), minimizing vertical and lateral physical forces is key for scrutinizing biological samples reliably at a high resolution. It is therefore of utmost importance to understand the different interaction mechanisms between the tip and the surface. Consequently, the selection of suitable cantilevers, imaging procedures, and the optimization of regulating parameters (e.g., feedback loop, scanning

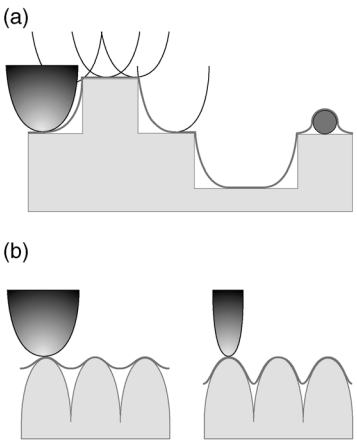


Figure 20.3 Superposition effects between the tip and the sample can distort the AFM topography. (a) Owing to its finite diameter ($\geq$10 nm in most cases), a cantilever tip cannot trace very sharp edges. Therefore, the imaged topography (gray line) is a superposition between the tip and the sample surface. (b) In many cases, however, the periodicity of the biological structures can be resolved correctly irrespective of the tip diameter.

Corrugation waviness of a surface.

speed, image size) are key experimental parameters that require meticulous tuning. For example, the type of ions (i.e., monovalent or divalent) and their concentrations play a decisive role in increasing the resolution of a map; they are to be selected such that a repulsive (≈0.05 nN) and long-range (several nm) interaction is generated between the tip and the object. This strategy increases the contact force marginally with only a small fraction acting locally on the protein structures thus preventing sample deformation in most cases. However, this tuning of the interactions is associated with an increase in the gross force of the coupled cantilever–sample system, further suppressing the natural thermal resonance of the cantilever spring (Section 20.5). Advantageously, this reduction in the thermal noise increases the signal-to-noise ratio, resulting in a high resolution of the AFM topography.

20.4 Preparation Procedures

Because the general working principle of the AFM is based on sensing and detecting molecular interaction forces, it is not necessary to metal coat or fluorescently label biological macromolecules or cells to identify them. The most important and often the only sample preparation step required is to immobilize the biological sample on an atomically flat support. This is a mandatory prerequisite for high-resolution imaging because the corrugation of an atomically flat surface hardly superimposes with the morphology of the soft biological sample and enables precise spatial control of the cantilever tip on the specimen. The sample preparation therefore requires finding a balance between the need to anchor the biological object while at the same time minimizing the interactions between the object and the support surface to ensure the mapping of an unaltered, native state (conformation) of the sample (protein).

Sample supports used in optical and electron microscopy are also utilized for AFM applications. Examples include glass, muscovite mica, graphite, and metal (such as gold-coated) surfaces. Based on the desired application, each of these sample carriers can be tuned to have unique physical and chemical properties (e.g., surface charge or roughness). For example, mica, which is characterized by a layered crystalline structure, is ideally suited for the immobilization of proteins and nucleic acids. An adhesive tape can be used to peel off crystalline layers from the underlying surface to provide a relatively chemically inert, negatively charged, and an atomically flat surface.

The most commonly applied immobilization strategies are based either on physical interactions between the biological sample and a chemically inert surface or on the covalent attachment of the sample to a functionalized (reactive) support. Physical adhesion is mainly achieved by shielding the repulsive electrostatic interactions between the biological object and the sample carrier. Thus, increasing the ion concentration leads to an attractive interaction between the sample and the support surface, which is sufficient to immobilize the biomolecules. For this purpose, the buffer solution containing the biological sample is applied directly onto a freshly split mica surface, the macromolecules allowed to adsorb firmly for a few minutes, and their surfaces mapped with a cantilever tip. For chemical coupling, the sample support is first functionalized with a chemical moiety, for example, glass with silanes or gold with thioalkanes, and in a final step the biomolecules are reacted with the chemically active surface.

Because biological structures require water molecules for their structure–function dynamics, drying biological samples and imaging in air should be avoided as far as possible. During the drying process, biological molecules are exposed to tensile forces generated by the surface tension of water often leading to the collapse of the samples. Irrespective of the fact that drying artifacts can be minimized or avoided by certain vacuum sublimation procedures, biological samples, whenever possible, should be prepared and imaged in aqueous solutions to preserve their native structural and functional integrity.

20.5 Mapping Biological Macromolecules

Using the AFM, surface structures and dynamic processes of a range of different biological samples can be observed under native conditions. These include biological macropolymers (e.g., proteins, DNA, RNA, and polysaccharides), supramolecular complexes (e.g., metaphase

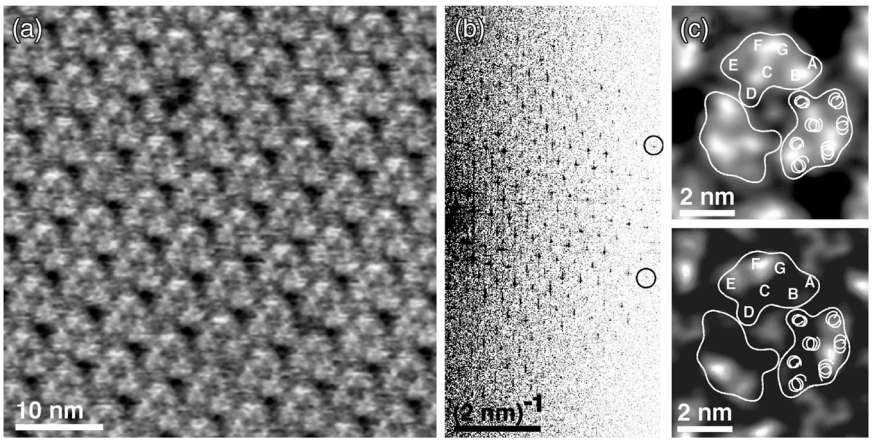


Figure 20.4 Cytoplasmic surface of the native purple membrane of *Halobacterium salinarum*. (a) The AFM topograph clearly reveals the assembly of single bacteriorhodopsin molecules into trimers; the trimeric units are further organized in a hexagonal lattice. The membrane topography was recorded in a physiological buffer at room temperature. (b) The diffraction pattern of the topography extends up to the 11th order (drawn circles), which suggests a lateral resolution of 0.49 nm. (c) The average topography (top) and the corresponding standard deviation (bottom) of the bacteriorhodopsin trimer allows correlation to structural data obtained by electron crystallography. Superimposed is the outline of bacteriorhodopsin molecules, as well as the positions of the seven transmembrane α-helices (A–G).

chromosomes), as well as bacterial or cellular associations, and tissues of higher organisms. Currently, the highest resolution is achievable on isolated macromolecules immobilized on an atomically flat surface. The technique is capable of providing high-resolution topography of individual membrane proteins in their native lipid bilayer environment, while at the same time the resolution is high enough to monitor and discern subunit dynamics such as associated with helix and polypeptide loop movements during the opening and closure of membrane channels. Figure 20.4 shows the cytoplasmic surface of the purple membrane of the archaebacterium *Halobacterium salinarum*. This purple membrane consists of lipids and the seven transmembrane α-helical protein bacteriorhodopsin, a light-driven proton pump. The AFM topography of the native membrane surface clearly shows the natural crystalline arrangement of bacteriorhodopsin – individual bacteriorhodopsin molecules assemble as trimers, which further form a two-dimensional hexagonal lattice. The close packing allows the maximum density of proton pumps for light-driven energy production. The AFM topography evidently portrays the molecular (polypeptide loops and termini) variability of each bacteriorhodopsin assembly. Nevertheless, the statistical average of the individual particles reveals a representative structural snapshot of bacteriorhodopsin, and the calculated standard deviation is a measure of the conformational variation of the population. In addition, it is useful to compare the averaged AFM topography with the structural information obtained from complementary structural biology methods (e.g., X-ray crystallography, electron microscopy, or NMR). For example, superimposing the mean contour from AFM on the three-dimensional X-ray structure makes it possible to assign the surface details to the secondary structures, and to further study the structural dynamics of the protein under different conditions. (Figure 20.4)

X-Ray Structure Analysis, Chapter 21

Electron Microscopy, Chapter 19

NMR Spectroscopy of Biomolecules, Chapter 18

The AFM is also used to characterize the stoichiometry and the formation of membrane proteins in functional complexes. Figure 20.5a shows the Na^+-ion driven rotors of the F_OF_1-ATP synthase from *Ilyobacter tartaricus;* the number and arrangement of the subunits of the functional rotor can be unambiguously recognized. The high-resolution tool provided a straightforward approach to illustrate that the cation-fueled (H^+ or Na^+) F_OF_1-ATP synthases from different organisms are constructed of different numbers of subunits. An attractive although debatable speculation is whether the different ATP synthases evolved to adapt the number of rotor subunits to the membrane potential of the cell and thus regulate ATP synthesis for energy consumption. Thus, elucidating the cellular mechanism that controls rotor stoichiometry will advance our understanding of ATP synthesis.

AFM is also applied for observing dynamic processes such as DNA transcription, protein diffusion, conformational changes of individual proteins, and even the emergence of molecular networks. The time required to record an AFM topograph is a decisive factor in determining the temporal resolution of the dynamic process. Depending on the system being investigated and the AFM setup, the recording time is between 1 s and 15 min. The so-called "gap junctions" or the communication channels from the rat liver cells are shown in Figure 20.5b and c. The individual connexins of the gap junctions show a nearly perfect hexagonal packing. The central pores of the individual hexamers are clearly discernible in the unprocessed topographs. In the presence of a signaling molecule such as calcium in the buffer solution, reversible closing of the channels can be directly observed (Figure 20.5c). The average surface structures of the channels

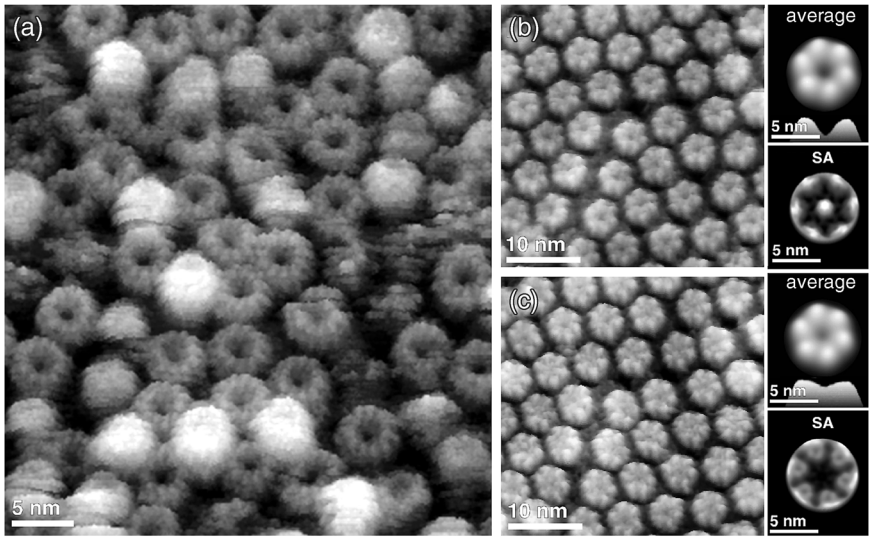


Figure 20.5 Determining the protein complex assembly and function by AFM. (a) Na^+-driven rotors of F_oF_1 ATP synthase from *Ilyobacter tartaricus;* the high resolution topography allows to characterize the stoichiometry and arrangement of the subunits of a functional rotor. (b) The extracellular surface of purified communication channels (gap junctions) from rat liver epithelial cells clearly shows hexameric proteins with an open central channel. The average and the standard deviation (SD) allow insights into the structure and its conformational flexibility. While the average profile reveals the channel entrance, the SD map assigns an increased structural flexibility to the central channel. (c) In the presence of 0.5 mM Ca^{2+} (at neutral pH) the central channel of the hexamer is closed. This mechanism is evident in the average hexamer structure, and the associated SD map depicts a loss in flexibility of the channel entrance in the closed state. All topographies were recorded in a physiological buffer at room temperature.

and their standard deviations provide insight into the structural variability and map the flexible regions of the protein. An open communication channel demonstrates structural variability and flexibility; these characteristics are lost upon channel closure. This observed relationship between flexibility and functional conformational change of the channel is not an exception, and has been determined previously for other proteins. Thus, to a large extent it can be generalized that the flexibility of structural features often correlates with their ability to perform functionally related conformational changes. The advent of the ultrafast AFM, capable of taking several hundred pictures per second, is a major breakthrough in AFM development, and will allow real-time observation of various dynamic processes under native conditions.

In contrast to the sub-nanometer imaging of individual molecules, a maximum resolution of 50 nm is attained for AFM imaging of cells and tissues. The low resolution is partially owing to the flexibility and dynamic motion of living cells, but also due to significant roughness of the cell surface. Although the AFM still provides important physical insight into cellular function, the low resolution of cell surface imaging hinders identification of the observed structures on cells. To overcome this shortcoming, AFMs are often combined with modern light microscopy techniques. The combination makes it possible to correlate topographic information of a cell with its superficial structures (e.g., vesicles or the cytoskeleton, which need to be fluorescently labeled for simultaneous AFM and optical microscopy). In addition to the determination of structural features, the AFM permits the determination of various other physical parameters (e.g., the elasticity of a biological object). The cantilever spring deflects when the tip is pushed against a sample; this signal is analyzed to characterize the elastic properties of the sample. Importantly, an unambiguous interpretation of cell elasticity maps requires the identification of different structural components that contribute to the mechanical stability of the cell. From a morphological standpoint, the cytoskeleton thereby plays a crucial role, can be directly traced by tip-induced deformation of the flexible cell membrane. Alternatively, it is also useful to analyze the phase changes of an oscillating cantilever spring during the dynamic mapping of a cell. Phase modulation reflects changes in the elastic properties, charge, and roughness of a cell surface.

20.6 Force Spectroscopy of Single Molecules

The ability of the AFM to detect forces with pN sensitivity helps to characterize the strength of biological and chemical bonds, and the behavior of individual molecules under mechanical stress. An attractive straightforward measurement is between ligands and receptors. This is achieved by first functionalizing the cantilever tip with the protein or the small molecule ligand, and then using the probe to interrogate the binding partner on a sample support. Using this approach it has been possible to detect specific binding and quantify molecular interactions

Table 20.2 Estimated rupture forces of chemical and biological bonds.

Bond	Rupture force (pN)[a)]
Covalent	1000–2000
Cell–cell interaction	500
Biotin–avidin	200
Antibody–antigen	<200
Protein (selectin)–sugar	100
Protein–protein or protein–DNA	10–50
Hydrogen bond	a few pN

a) At an approximate pulling speed (loading rate) of 500 nm s^{-1} (5 nN s^{-1}).

between proteins and ligands, proteins and nucleic acids, antibodies and antigens, and cells (Table 20.2). In these so-called force-spectroscopy experiments the deflection of the cantilever spring is measured while the functionalized tip (e.g., biotin) is retracted from a surface functionalized with the complementary binding partner (e.g., avidin). As a result of the cantilever retraction and consequently its continuous bending, a slow incremental force is exerted on the bond. The structural transition, that is, the rupture of the interaction, is characterized by a sudden change in the spring deflection. The interaction force depends on the pulling speed and, in most cases, is only a few tens to hundreds pN (Table 20.2). These forces are a result of the collective molecular interactions that contribute to the biological mechanism. The dynamic range of the rupture forces determined at different velocities gives insight into the nature of the energy landscape (Figure 20.6). The force spectroscopy approach has also been applied to estimate the elasticity of single polysaccharides and nucleic acid chains, controlled unfolding and refolding of individual proteins.

The first proof-of-principle mechanical unfolding experiments on the water-soluble modular muscle protein titin and the transmembrane protein bacteriorhodopsin showed that the AFM is a versatile, sensitive tool with capabilities to quantify the mechanical forces associated with the global denaturation of a single protein as in the case of titin, and to deconvolute individual secondary structure elements of membrane proteins with a high resolution. In contrast, only the average properties of a molecular ensemble are determined in conventional chemical or thermal (un)folding experiments. Thus, the power of AFM extends beyond precise spatial manipulation to dynamically follow temporal fluctuations of single molecules that are lost in a sea of molecules in bulk experiments. Furthermore, force spectroscopy can also assist in investigating

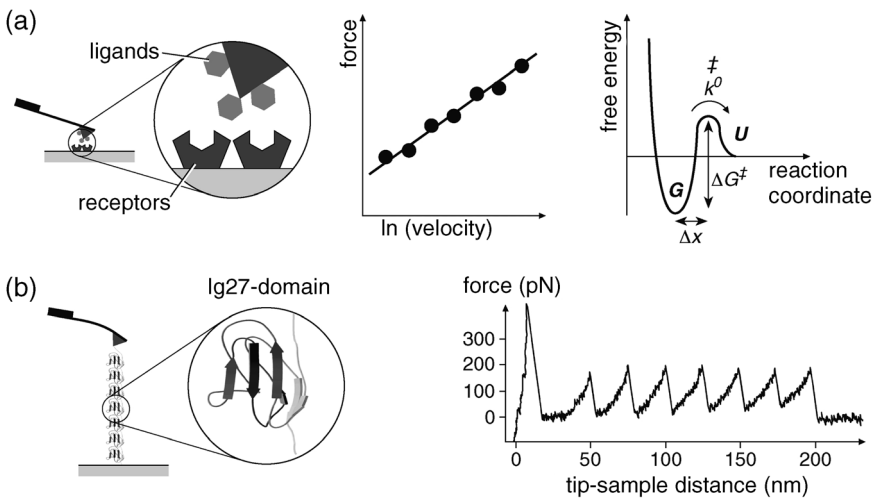


Figure 20.6 Force measurements on single molecules. (a) The AFM is used to determine the interaction forces between individual cells or molecules. For example, the cantilever tip is functionalized with ligands and brought in contact with the interacting receptors bound to a sample carrier. The interaction forces of the molecular bonds are measured by separating both surfaces at a certain velocity; the force increases with increasing separation speed. The kinetics of the forced separation can be described by a two-state energy landscape with the bound state (*G*) separated from the unbound (*U*) state by an energy barrier; $\Delta G^‡$ is the barrier activation energy between *G* and *U*, k^0 is the natural transition rate, and Δx is the width of the potential barrier. Both k^0 and Δx can be determined from the velocity dependence of the rupture force. (b) A multi-domain protein construct (e.g., immunoglobulin-27) can be stretched with the AFM, and the unfolding force of each domain can be quantified. The sequential unfolding of single domains results in the characteristic saw-tooth pattern of a force–distance curve.

the effects of point mutations, temperature, buffer solutions, and ligands or analytes on the stability and folding kinetics of folding and unfolding intermediates of water-soluble and membrane proteins.

20.7 Detection of Functional States and Interactions of Individual Proteins

Exerting a sufficient pulling force at a terminal end of a membrane protein triggers a gradual and sequential unfolding of its secondary structure elements until the whole protein is extracted from the lipid bilayer membrane. As mentioned above, the unfolding force of each structural segment is a measure of its stability, which is influenced by the various inter- and intramolecular interactions. In principle, these interactions can be altered by the binding of a small molecule or a ligand. It was hypothesized that, if a sensitive enough cantilever is used, force spectroscopy should be able to detect the binding of a ligand or an analyte to a membrane receptor. In 2005, experimental evidence of the hypothesis was provided with a convincing demonstration of Na^+ ion binding to the protein antiporter NhaA. The unfolding signals of single NhaA molecules contained the signature peaks denoting the precise location of the Na^+ binding site as well as the strength of the altered interactions resulting from ligand binding (Figure 20.7). Similarly, it was also possible to resolve the site-specific differences (and the corresponding changes in the molecular interactions) between Na^+ binding (the natural ligand) and an inhibitor of NhaA. In principle, the method could be further refined so that it is possible to distinguish between the binding of different ligands and drugs, ideally at the same putative site.

Other exemplary experiments have employed force spectroscopy to determine the functional states of a protein. In stark contrast to other methods, force spectroscopy bypasses approaches requiring labeling with dyes, molecular markers, or chemical and genetic modification of the

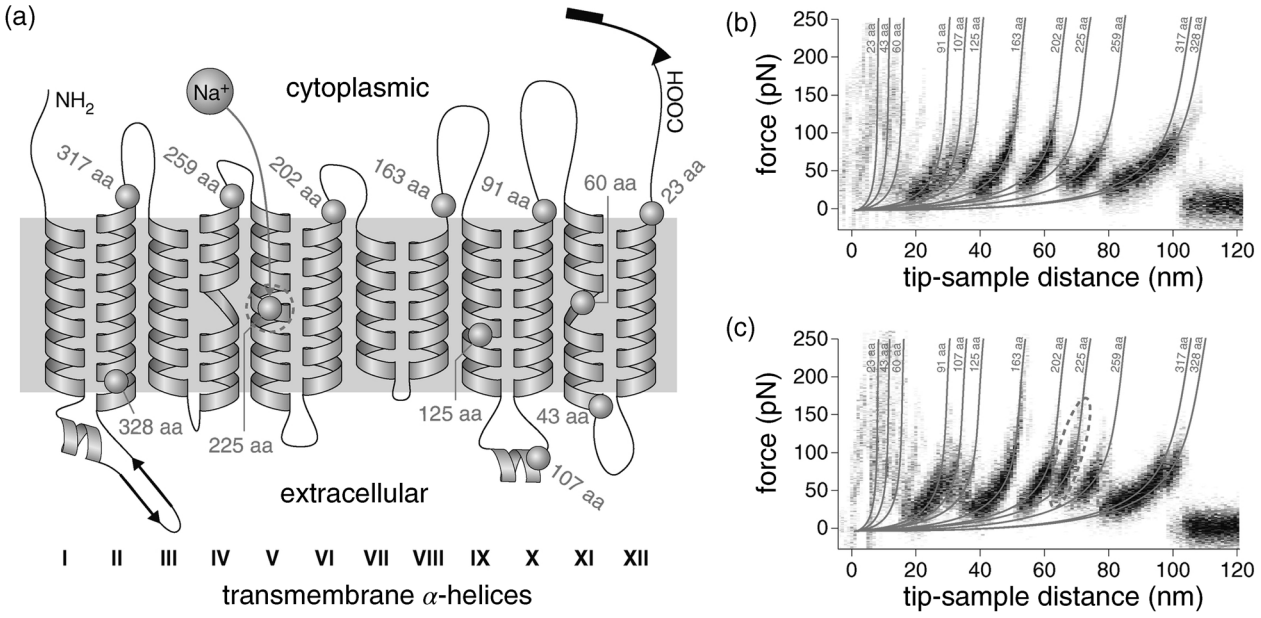


Figure 20.7 Detection of ligand binding and function of single membrane proteins by force spectroscopy. (a) The secondary structure of the sodium/proton antiporter (NhaA) from *Escherichia coli* is shown. The C-terminal end of a single antiporter is grabbed using the AFM tip and subjected to mechanical traction. At a high enough tensile force (≈200 pN) the secondary structure elements of the membrane protein gradually unfold. (b) The individual peaks of the force–distance curves quantify the interactions that stabilize the secondary structure elements (gray circles in (a)). To localize these interactions within the membrane protein, the force peaks are approximated using the worm like chain (WLC) model (gray curves in (b) and (c)). The WLC fits to the force peaks provide a means to determine the tip–sample distance, which can be converted to obtain the lengths of the stretched polypeptide chains (in nm or amino acids). The height of a force peak quantitates the interaction forces of the bonds. (c) While the force–distance curve in (b) resulted from unfolding of a single NhaA without the ligand, the curve in (c) shows NhaA unfolding in the presence of its native ligand (Na^+ ion). On the basis of a new force peak (highlighted by gray dashed ellipse) it can be suggested that ligand binding introduced an interaction. On the basis of the force peak, the new interaction can be assigned at amino acid position 225 (counted from the C-terminal end) of the secondary structure.

membrane protein, but the technique rather relies on the interaction between the cantilever tip and the native membrane protein, and a chemical agent triggering conformational remodeling. These examples support AFM-based force spectroscopy as a technique well on its way in establishing itself for the mechanical–structural–functional analysis of individual proteins. Currently, automated devices are being developed that can perform experiments on single molecules and analyze the data with minimal user input. Thus, it is highly likely that force spectroscopy will find a stronghold in routine industrial and bioanalytical applications in the foreseeable future.

Further Reading

Dokland, T., Hutmacher, D.W., Ng, M.-L., and Schantz, J.-T. (eds) (2006) *Techniques in Microscopy for Biomedical Applications*, World Scientific Press, Singapore.

Rigler, R. and Vogel, H. (eds) (2007) *Single Molecules and Nanotechnology*, Springer, Heidelberg.

Gerber, C., and Long, H.P. (2006) How the doors of the nanoworld were opened. *Nat. Nanotechnol.*, **1**, 3–5.

Muller, D.J. and Dufrêne, Y.F. (2008) Atomic force microscopy as a multifunctional molecular toolbox in nanobiotechnology. *Nat. Nanotechnol.*, **3**, 261–269.

X-Ray Structure Analysis

21

Achim Dickmanns and Ralf Ficner

Georg August Universität Göttingen, Institute for Microbiology & Genetics (GZMB), Department for Molecular Structural Biology, Center for Molecular Biosciences, Justus-von-Liebig Weg 11, 37077 Göttingen, Germany

Knowledge of the three-dimensional structures of biological macromolecules has a significant impact on our understanding of their cellular function. X-Ray crystallography as well as the already mentioned nuclear magnetic resonance spectroscopy (NMR, Section 18.1) and electron microscopy (EM, Chapter 19) are the most important methods for structure determination in molecular biology. To date (July 2014) more than 100 000 three-dimensional structures of biological macromolecules have been deposited in the Protein Data Bank (PDB), 88% of which are crystal structures, 11% are NMR structures, and 0.7% are EM structures.

Besides X-ray crystallography, which allows structure determination at ångstrøm (Å) resolution (10^{-10} m), an additional X-ray based method is available, namely, small angle X-ray scattering (SAXS), which allows structure determination in solution at a lower resolution. An emerging method of structure determination uses a free electron laser (FEL) for diffraction imaging of microcrystals, but is aimed at structure determination of individual molecules in solution. The latter two methods will be discussed in more detail at the end of this section.

The methods for structure determination by means of X-rays differ not only in the achievable resolution range correlating with structural details (Figure 21.1) but also in the way the sample is prepared and how the acquired data is analyzed (Figure 21.2).

In crystallography, the purified macromolecule (protein, DNA, RNA) or macromolecular complex has to be brought into a crystalline state. Nano-crystals are sufficient for the FEL, whereas soluble macromolecules are needed for structure determination using SAXS.

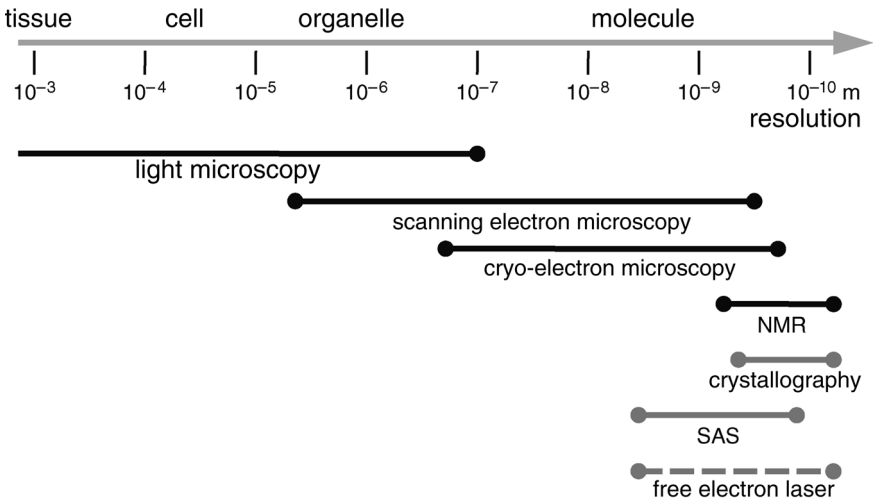


Figure 21.1 Resolution range of the different methods for structure determination of biological macromolecules.

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

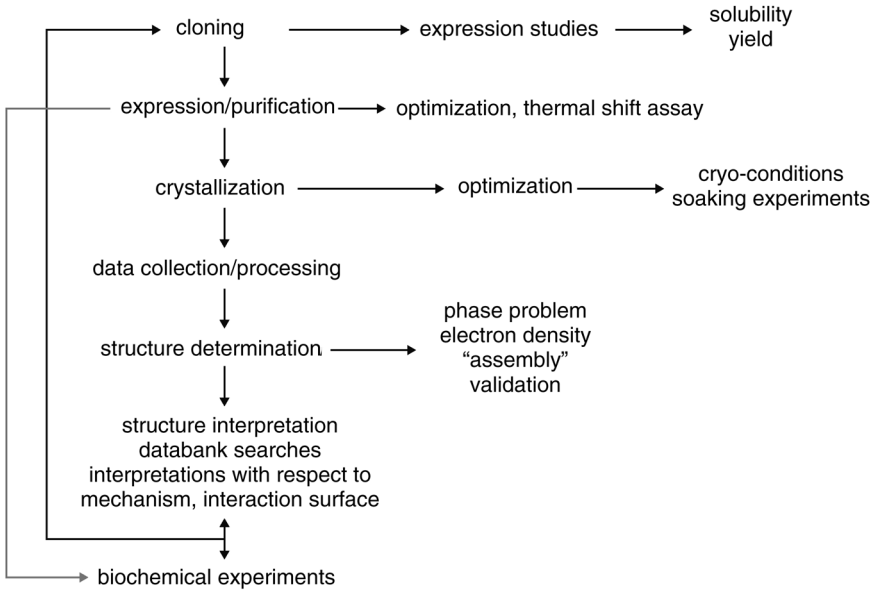


Figure 21.2 Modus operandi and working steps in X-ray crystallography.

To date, X-ray crystallography is still the method of choice to obtain high-resolution structures, but the recent developments using FEL have led to significant progress in the structure determination from micro-crystals that are not suited for standard X-ray diffraction experiments. A similar resolution may be obtained by NMR spectroscopy, which has the advantage of using protein in solution, thereby avoiding the difficult and chance-dependent task of crystallization. The caveat still in NMR spectroscopy is the difficulties in determination of structures with a molecular weight exceeding 40 kDa, whereas crystallographic methods allow the investigation of proteins, protein complexes, and even whole assemblies like ribosomes with the overall molecular mass exceeding 1 MDa.

21.1 X-Ray Crystallography

X-ray crystallographic methods allow the three-dimensional structure determination of proteins, RNA, DNA, and macromolecular complexes thereof at atomic resolution, resulting in a structural model with the precise determined positions of the individual atoms. Besides the information about the overall fold and secondary structure elements, the structure allows insight into the protein function at an atomic level, such as the catalytic mechanism of an enzyme or the specific interaction of a transcription factor to DNA.

In principle, biological macromolecules of any kind and size may be investigated – as long as they form diffracting crystals. The first crystals of proteins were described even before the discovery of X-rays, like those of the light harvesting protein phycoerythrin from red algae in 1894. Ground laying work on salt crystals by von Laue and Bragg in the first decades of the twentieth century allowed the use of beams scattered on a crystal to deduce a picture of the arrangement of the individual atoms within the crystal. The mere size of the protein crystals required novel crystallographic methods, which were used for the first time by Perutz and Kendrew for the structure determination of hemoglobin and myoglobin, respectively.

Since the first successful protein structure determination by means of X-ray crystallography in the late-1950s almost 90 000 structures have been determined using this method, encompassing membrane proteins, protein–nucleic acid complexes, large oligomeric protein complexes, and even intact viruses.

A structure determination by X-ray crystallography requires several working steps that will be presented in the following in more detail: crystallization, characterization of the crystals, recording of diffraction data, phase determination, interpretation of an electron density map as well as refinement of the structural model. The quality of the obtained crystal structure is described by various criteria, which will be commented on subsequently.

21.1.1 Crystallization

For X-ray diffraction experiments the macromolecule of interest is required in a crystalline state in the form of singe crystals with dimensions ranging from a few to hundreds of micrometers. Any kinds of crystal pathologies (several crystals growing together resulting in overlapping lattices, twinning, long-range disorder) usually disqualify crystals from being used for structure determination.

Many of the time-consuming steps in the first decades of protein crystallography, like collection of diffraction images and their processing or refinement of the structural model, have been dramatically shortened due to technical developments and increased computing power. Despite these developments, the crystallization of biological macromolecules is still a technique based on trial and error, thus evading prediction of the best crystallization conditions.

The individual properties of proteins, like molecular weight, amino acid sequence, or isoelectric point do not allow the deduction of conditions that would be more or less suitable for crystallization. Due to the improvement in overexpression and subsequent purification it is nowadays possible to obtain large amounts of purified protein that together with automatization of the crystallization process allows testing of thousands of different crystallization conditions for various constructs of the protein of interest in a short time period.

For crystallization trials as well as small angle X-ray scattering experiments the protein has to be purified as close as possible to homogeneity. The purity and homogeneity are usually determined by SDS and IEF (isoelectric focusing)-polyacrylamide gel electrophoreses. More sophisticated methods are size exclusion chromatography and "in line" or separately performed dynamic light scattering. A quite recent method for determination of protein stability in a certain buffer is the so-called "thermal shift assay," which is based on the temperature dependent unfolding of the protein. The unfolding is detected via the binding properties of a fluorophore (Figure 21.3).

Electrophoretic Separation Techniques, Chapters 11, 12

Chromatographic Separation Techniques, Chapter 10

Spectroscopy, Chapter 7

The actual process of crystallization of macromolecules is driven by supersaturation, which controls the rate of simultaneously occurring processes of nucleation and crystal growth. Depending on the conditions achieved by addition of a so-called precipitant, one of these two processes could be dominant over the other and results in crystals of different shapes, sizes, and quantity (Figure 21.4). During the actual process of crystallization the macromolecules move from the supersaturated solution into a solid state. This is generally achieved by addition of the precipitant, which supersedes the macromolecules out of the solution (Figure 21.4). Thereby, the precipitant concentration is intended to increase slowly either by diffusion of water out of the protein solution in the mother liquor or by diffusion of precipitant into the protein solution.

The most widely used experimental setups are the *hanging drop* and the *sitting drop*, which use vapor diffusion to slowly increase the concentrations of macromolecule and precipitant in the crystallization droplet. For crystallization by the hanging drop method a small volume (1–2 µl) of a highly concentrated protein solution (5–40 mg ml^{-1}) is placed on a coverslip

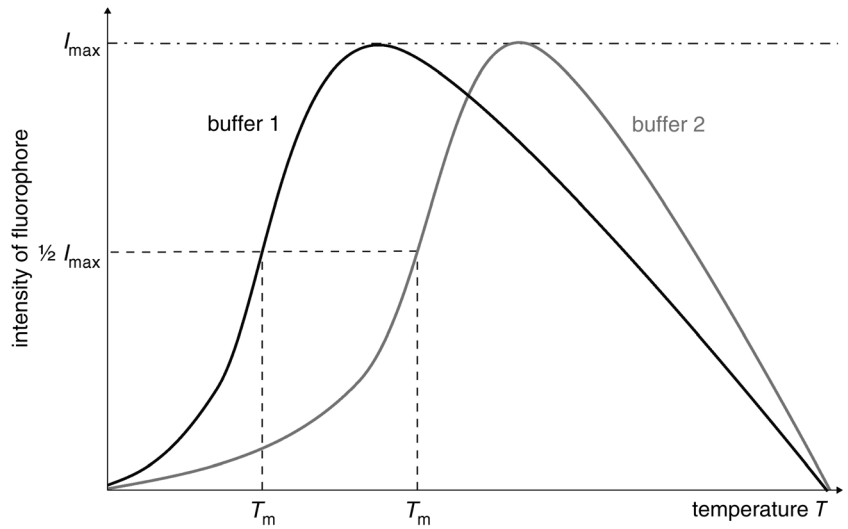


Figure 21.3 The thermal shift assay is a method for determining the dependence of protein stability on the buffer solution using thermal denaturation. Due to the temperature induced unfolding of the protein a fluorescent dye binds to the protein, thereby changing its fluorescent properties. This change is detected and plotted against the temperature. At a certain temperature, which is specific for the protein and buffer, the denatured proteins start to aggregate, thereby releasing the dye, leading to a decrease of the intensity measured. In an ideal case, the half-maximal change of the intensity during the initial unfolding correlates to a specific temperature, which is used as a marker for the stability of the protein. Note that stability and functionality must not necessarily be optimal in the same buffer condition.

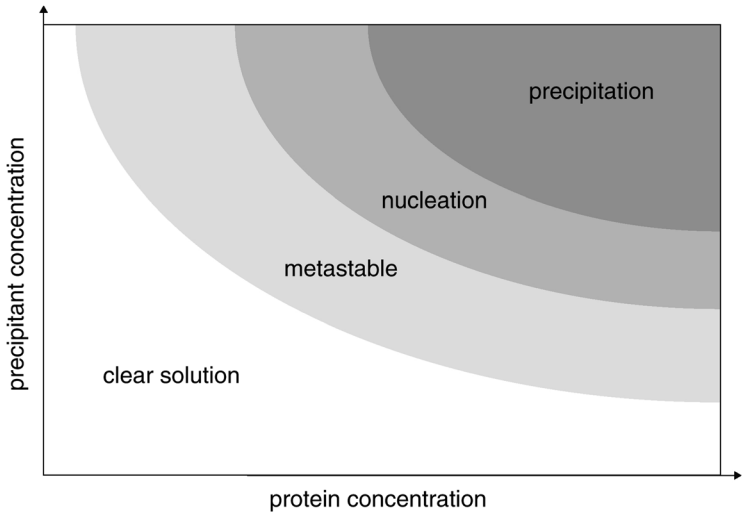


Figure 21.4 Phase diagram of an aqueous protein solution showing zones for crystal nucleation, growth, and precipitation. Crystal formation is a process that may be roughly divided into three steps. For a successful crystallization a state within the nucleation zone has to be reached. The crystal nuclei then grow in the metastable region into crystals that can be used for measurements. The crystals grow only in a supersaturated solution, once supersaturation is exhausted the crystals stop growing. In the region of precipitation too much protein is present, whereas a clear solution indicates that the protein and/or precipitating agent concentrations are too low.

(Figure 21.5). Subsequently, the same volume of the reservoir solution or mother liquor (the mixture of precipitant and buffer) is added. After mixing the drop should remain clear. In the case of immediate turbidity, the mixing should be repeated and either the protein or precipitant concentration should be reduced. The coverslip is immersed and place over a well, containing 0.5–1.0 ml of the undiluted mother liquor, and its rim covered by a thin layer of silicon oil or fat, leading to a sealed, closed system (Figure 21.5). Because the precipitant concentration in the *hanging drop* is only half of that in the reservoir, vapor diffusion causes a reduction of the hanging drop volume. This, in turn, leads to a slow increase of the concentration of protein and precipitant in the *hanging drop,* until oversaturation is reached and crystallization of the protein might occur.

The *sitting drop* method is based on the same principle, but has the advantage that the system can be sealed by clear tape (Figure 21.6). Other, but more rarely used, methods for crystallization of macromolecules are micro-dialysis and crystallization in a gel matrix or under oil. These methods are described in more detail in the books cited in the Further Reading.

The most widely used precipitants are salts (like ammonium sulfate, sodium citrate, sodium potassium phosphate) and organic compounds like poly(ethylene glycol) (100–20 000 Da) or alcohols (methylpentane-diol, ethanol, isopropanol).

Moreover a combination of various precipitants or the addition of other salts (e.g., magnesium sulfate, calcium chloride) or the use of detergents could be tested. In classical crystallization trials, the grid screens, parameters like precipitant concentration are systematically varied, allowing the testing of a high number of possible crystallization conditions. An alternative is the sparse-matrix screens, which are based on the statistical evaluation of

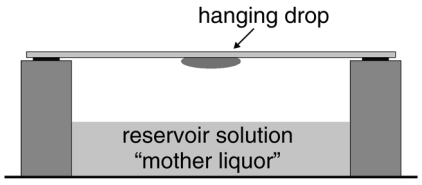


Figure 21.5 Protein crystallization using the *hanging drop* method. On a cover slip hangs a drop of 2–6 µl volume, containing protein, precipitant, and buffer. At the bottom of the well is 1 ml of a precipitant–buffer solution of higher concentration than in the droplet. By the vapor diffusion of water molecules the concentration gradient between the drop and the reservoir is reduced till equilibrium is reached. Thereby, the drop volume is reduced and the solubility barrier of the protein is crossed and the protein either precipitates or reaches a supersaturated state needed for nucleation.

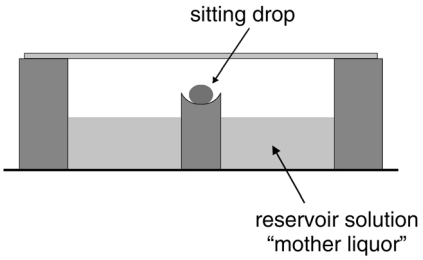


Figure 21.6 Protein crystallization using the *sitting drop* method. The mixture of protein sample and mother liquor is placed in the center of the elevated well, surrounded by the mother liquor. The process of crystallization is based on vapor diffusion, as in the hanging drop method.

successful crystallization experiments. Conditions that more frequently lead to successful crystallization may be grouped in a single screen, but neighboring conditions do not necessarily need to be correlated with respect to composition. The choice of conditions may be adapted to the properties and stability of the macromolecule to be crystallized (e.g., for protein–RNA complexes or multi-protein complexes). An important role is played by the actual pH under the crystallization conditions; it is usually varied in small intervals, ranging from pH 4 to 10.

Because crystallization is strongly dependent on temperature, various incubation temperatures should be used in parallel (e.g., 4 and 20 °C). The number of conditions that can be tested in a short period of time has been dramatically increased by using the sophisticated pipetting robots, allowing thousands of crystallization conditions to be tested. Modern automated crystallization systems are made up of several components: a pipetting robot that sets up the various crystallization conditions from stock solutions, a nanoliter pipetting robot transferring the crystallization conditions in 96-well plates. Moreover, it varies the crystallization conditions and the protein solution in small droplets, mostly ranging between 200 and 400 nl in volume. The final component is a system for storage of the plates connected to an automated imaging system that takes pictures at preset time points, allowing changes in the crystallization droplets to be tracked over time.

For some proteins single crystals grow out of the initial screens that diffract well and may be used directly for structure determination (Figure 21.7). In most of the cases it takes weeks or months to achieve this goal. Normally, initial tests yield – if at all – small crystals (microcrystalline precipitate, micro-platelets, or needles), often attached to each other, so that their growth conditions have to be optimized to obtain single crystals suitable for diffraction tests.

In a first round of optimization, small changes in the crystallization solution, the pH, and/or the precipitant concentration are made to obtain bigger single crystals. Subsequently the best condition is optimized further by addition of so-called additives or detergents, which might result in changes in the crystal packing leading to bigger and/or better diffracting crystals.

Detergents, Section 1.8

In addition, seeding techniques are used to increase the size of single crystals. Here, microcrystals are transferred from the original condition using a diluting loop or a cat whisker into a new drop of protein and crystallization solution and serve as starting points (seeds) for the growth of crystals.

Sometimes the resulting crystals are optically perfect single crystals, but exhibit a weak and low resolution diffraction, resulting from lack of long-range order due to misarrangement of the macromolecules in the crystal lattice. The long-range order may be improved by addition of other salts or organic molecules (so-called additives). An alternative possibility is the controlled reduction of the solvent content within the crystal, for example, by using a commercially available free mounting system or crystal humidifier.

The frequently asked question about the minimal protein amount required for a crystallographic structure determination cannot be answered generally. An initial screen usually consists of 500–1000 different conditions, which would require 5–10 mg of pure protein if 1 µl of protein solution is used per condition at a concentration of 10 mg ml^{-1}. The use of a fully automated nanoliter robot would reduce this amount by a factor of 5–10.

If the macromolecule is a mutant of a protein whose 3D structure has been already determined or is homologous to proteins with known 3D structure, the whole structure determination process could be performed without the need to prepare derivatized protein or perform soaking experiments. In lucky cases, only one crystal could be sufficient, because the so-called "phase problem" (Section 21.1.3) may be solved with the help of the molecular replacement method, which requires the known structure to be used as the starting model. If the macromolecule has an unknown fold, many crystals are required in order to find suitable heavy atom derivatives that help to solve the phase problem. For such a *de novo* structure determination, on average 1–50 mg of macromolecule is required. Alternatively, if a selenomethionine-containing variant of the protein can be produced and crystallized, only one crystal might be sufficient to solve the phase problem.

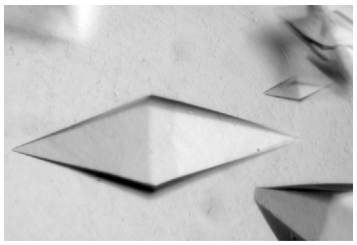


Figure 21.7 Single crystals of amylase C from *Thermotoga maritima*. The crystals belong to the space group $I4_122$.

21.1.2 Crystals and X-Ray Diffraction

Protein crystals may differ in size and shape, but generally crystals with an edge length of 0.02–0.3 mm are suitable for X-ray structure analysis. Crystals are composed of a huge number

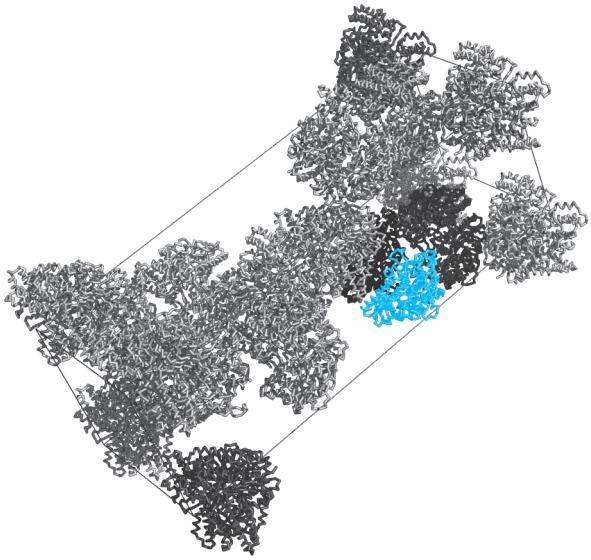


Figure 21.8 Crystallographic unit cell and crystal packing of the protein AmyC from *Thermotoga maritima*. The protein molecule in the asymmetric part of the unit cell is colored in blue. The position and arrangement of all other molecules within the unit cell are determined by the symmetry operations defined by the crystal's space group ($I4_122$).

of unit cells, which represent the smallest parallelepiped that by translation along its edges in all directions forms the complete crystal lattice (Figure 21.8). In the case of protein crystals the dimensions of unit cells range typically from ca. 20 to 2000 Å (1 Å = 10^{-10} m). Besides the primitive crystal lattice (P) variants like face or body or rhombohedral centered lattices also exist (C, F, I, and R) (Figure 21.8). The length of the cell axis and the angles between them are the cell constants of a unit cell.

With the exception of the triclinic system, in all other crystal systems the unit cells contain at least two asymmetric units related by primitive symmetry operation(s) defined by the crystal lattice. For protein crystals those symmetry elements are axis and screw axis, which combine rotation with translation along the axis. The combination of the 14 Bravais lattice types with the 32 point groups of symmetry results in various space groups. Theoretically, 230 different space groups exist, but in protein crystallography only 65 space groups occur, due to the existence of chiral centers in the amino acids of the proteins. Therefore space groups with symmetry operators like mirror planes or inversions are impossible, reducing the number of possible space groups (Figure 21.9). A more detailed description of crystal symmetry and space groups can be found in the literature cited at the end of this chapter (Further Reading).

After successful crystallization an initial characterization of the crystals takes place. This requires placing the crystal in a focused X-ray beam using one of the two different mounting techniques (loops, capillaries). Besides the resolution limit, the space group (crystal symmetry) and dimensions of the unit cell are determined. One has to keep in mind that protein crystals often have wide channels between the protein molecules, channels filled with buffer solution (usually around 50–60% of the crystal volume). This makes protein crystals quite fragile against mechanical strain and not resistant to changes of humidity. Therefore protein crystals should be surrounded permanently by a solution, in most cases the reservoir solution of the respective crystallization experiment. To adjust the crystal in the X-ray beam the crystal is either transferred into a small glass capillary with a diameter of 0.3–1.0 mm (the surrounding solution is gently removed and the ends are sealed with wax) or the crystal is flash cooled in a thin film of liquid held by a small nylon loop. In the latter case the crystal is then kept at a constant temperature of 100 K, which helps to minimize X-ray induced damage of the crystal and thermal motion of macromolecules.

The experimental setup for X-ray diffraction experiments requires an X-ray source, a mirror system for focusing of the beam, and a detection unit for the diffracted X-rays. In the laboratory, the X-ray beam is usually generated by an X-ray generator containing a rotating anode or sealed tube. The wavelength of the X-rays depends on the anode metal, for example, for the usual copper based sources it is 1.5418 Å, (Cu-K_α X-rays). However, the preferred X-ray sources are synchrotrons, which are part of large research facilities accommodating many beamlines dedicated to different research areas. Synchrotron radiation has the advantage of a high

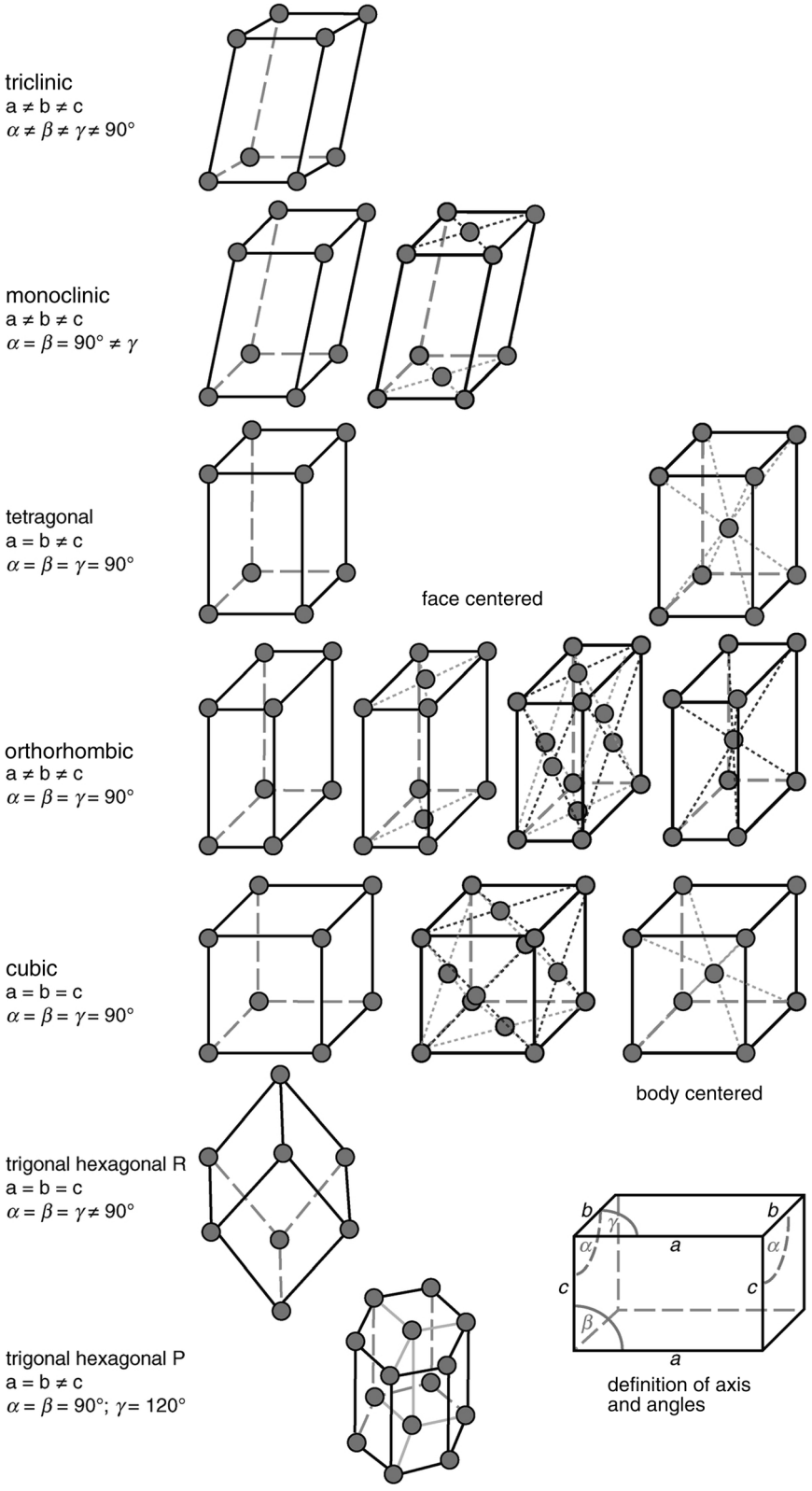


Figure 21.9 Fourteen Bravais lattices exist. There are four types of unit cells: primitive (P), side and face centered (A/B/C and F) as well as space/body centered (I) lattices. Together with the seven crystal systems (triclinic, monoclinic, orthorhombic, trigonal, tetragonal, hexagonal, and cubic) they result in the 14 possible Bravais lattice types.

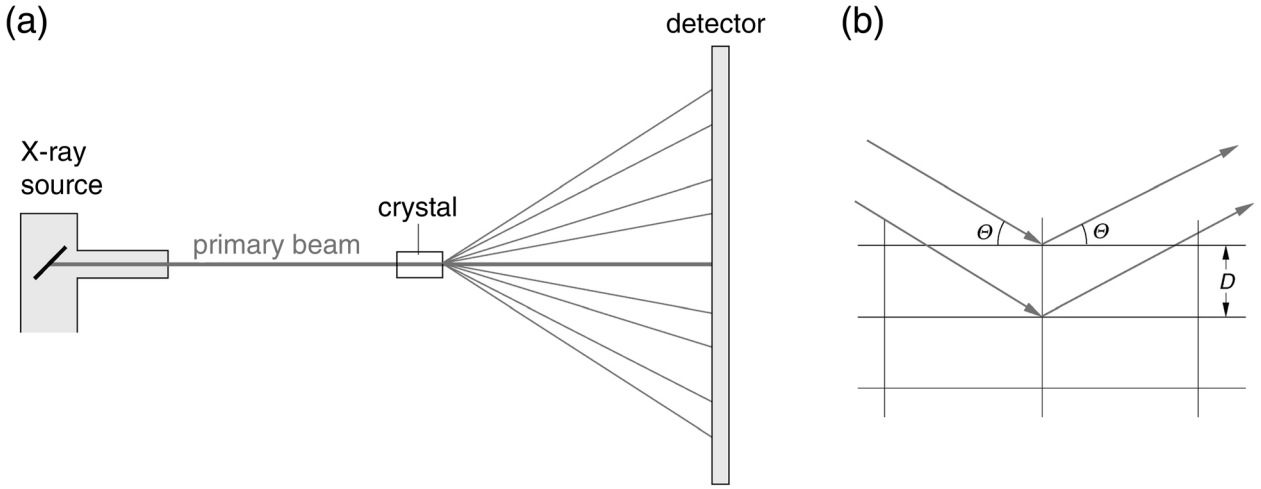


Figure 21.10 Basics of X-ray diffraction. (a) Schematic representation of an X-ray diffraction experiment on a crystal. The major part of the beam passes through the crystal directly; only a small portion is diffracted by the crystal. (b) Bragg's law describes the conditions for constructive interference at a set of parallel planes separated by the distance D. The reflection angle Θ depends on D and the wavelength λ.

brilliance and intensity of the beam and thus enables the use of smaller crystals for structure determination. Moreover, and in contrast to home source generators, the wavelength can be adjusted and set between 0.8 and 4.0 Å, allowing the measurement of datasets at multiple wavelengths as required for anomalous dispersion experiments (MAD, Section 21.1.3).

The basic physical principles for diffraction of X-rays on crystals apply for molecules composed of only a few atoms as well as macromolecules formed by tens of thousands of atoms. If an X-ray, also called the primary beam, hits the crystal most of the rays pass through the crystal without any interference. A small portion interacts with the electron hull of the atoms in the molecule and excites the electrons. When these excited electrons return to their ground state they emit X-rays in all directions. Since each individual atom is part of a repetitive array, interference of the emitted X-rays occurs. In most cases these interferences lead to extinction of the beam, but in some defined directions positive interference of the emitted beams takes place.

The diffraction of an X-ray beam by a crystal is described as reflection at a set of parallel planes by Bragg's law (Figure 21.10).

As noted, the diffraction of an X-ray beam with wavelength λ is described as a reflection at a set of parallel planes separated by the distance D. The value of the reflection angle Θ depends on D and λ.

$$2D \sin \Theta = \lambda \tag{21.1}$$

Since the arrangement of macromolecules in a crystal is not perfect and errors in the crystal lattice can occur, X-rays are diffracted only to a certain maximal angle Θ_{max}. According to Bragg's law the diffraction angle Θ_{max} correlates to the smallest detectable distance between two planes, d_{min}, which is defined as the resolution of the crystal structure:

$$d_{min} = \frac{\lambda}{2 \sin \Theta_{max}} \tag{21.2}$$

Thus, each individual spot on the detector, termed a reflex, corresponds to an X-ray diffracted at a specific cohort of planes. The position and orientation of the cohort of planes in the lattice are described by the Millers indices h, k, and l, which are defined as the intersection of the planes with the crystal axes. The values (h,k,l) correspond to the reciprocal coordinates of the intersections, with the lattice vectors set to the value of 1. For example, the plane dividing all lattice vectors in the middle (one half of the length) has the Miller indices (222).

Fourier Transformation, Sections 18.1, 19.4.2

In fact the diffraction of the primary beam is the result of the interference of all X-rays that are emitted from all the atoms within the crystal. The diffraction pattern contains data whose Fourier transform provides information on the contents of the unit cell.

To orient as many crystal planes as possible in a diffraction angle with respect to the primary beam the crystal has to be rotated during the measurement and diffraction images at different angles are collected. The strategy for data collection of diffraction data depends on the type and

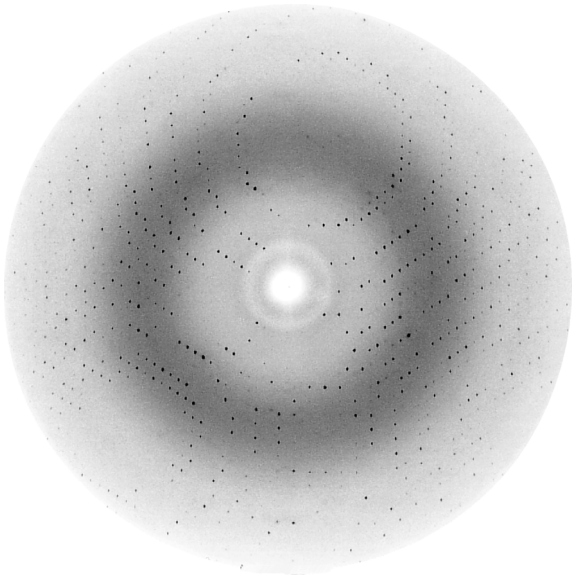


Figure 21.11 X-Ray diffraction image of a crystal of the RNA-modifying enzyme tRNA-guanine-transglycosylase. This enzyme modifies a nucleoside within the anti-codon of specific tRNAs by exchanging the base guanine with 7-aminomethyl-7-deazaguanine. The crystal was rotated by 1° during exposure and data collection. The spots at the rim of the image correspond to a resolution of 1.8 Å. Source: after: Romier, C. *et al.* (1996) *EMBO J.*, **15**, 2850–2857.

sensitivity of the detector used. X-ray sensitive films used in the early days have been completely replaced by area detectors that differ in the physical principles used to detect the X-rays and measure the intensity. The reasonably priced type of area detector for the laboratory is the so-called *image plate*, a plate of up to 35 cm in diameter with a thin layer of an X-ray sensitive material with barium-europium halogenides on its surface. At the site of impact the incoming X-ray shifts an electron from this material into a metastable state of higher energy. After excitation with a laser the metastable electrons fall back into the ground state, thereby emitting the energy released as short light pulses at a visible wavelength. This pulse of light can be read using a photomultiplier. The measured intensity is proportional to the intensity of the original X-ray. If an *image plate* detector is used the crystal is rotated with a constant speed (0.02–0.1° min^{-1}), usually by 0.2–0.5° per image. At the end of the detection, the exposure and rotation of the crystal are stopped, in order to readout the data. A typical diffraction image is shown in Figure 21.11.

An alternative to the image plate is a CCD detector, which has a very short read out time. The most recent developments of X-ray detectors are the CMOS detectors and the hybrid pixel array detectors. As with CCD detectors, a standard CMOS detector must first convert the X-ray photons into visible light using a phosphor layer on the surface of the detector. The visible light is then passed through a fiber optic stub to a photodiode of the CMOS circuit where analog to digital conversion takes place to give the data. In contrast, a hybrid pixel array detector directly counts X-ray photons. These most recent X-ray detectors perform data acquisition and readout in such short times that stops are no longer required, making them ideal for setups used at synchrotrons.

The total range of crystal rotation for a complete X-ray diffraction dataset depends on the space group of the crystal and the relative orientation of the crystal axis with respect to the primary beam and to the actual rotation axis of the crystal. The higher the symmetry of a crystal, the less it has to be rotated in total in order to obtain a full dataset. Usually data between 60° and 180° are collected also to obtain a certain degree of redundancy. The exposure time depends on the intensity of the X-ray beam but predominantly on the size and quality of the crystal. Using the rotating anode X-ray generator the exposure time per degree lies within 5–40 min, resulting in a complete dataset after 5–80 h. The high intensity of the X-ray beam at the synchrotron and the use of CCD detectors reduce this time to around 30 min per dataset down to a few minutes for the latest generation of detectors. The caveat of synchrotron radiation is the possibility of significant damage to the crystal during exposure, even though the crystals are kept at a temperature of 100 K during the X-ray diffraction measurement.

After data collection the diffraction images are used to determine the space group and the unit cell dimensions of the crystal with the help of data processing programs. Subsequently, the intensity of all detected reflexes is determined. Thereby, one has to be aware of the fact that

single reflections could be distributed over two or more subsequent images, making the scaling of the individual images to each other an important step in data processing. The quality of the X-ray diffraction data is described by *R*-factors like the R_{sym}. With exception of the triclinic space group defined reflexes are symmetry related and thus, theoretically, should have the same intensity. The deviation of the measured intensities of the symmetry related reflexes from this mean value is a good indicator of the quality of the measured data. For low and medium resolution data the R_{sym} is 3–5% off, and with increasing resolution the R_{sym} increases, because the intensity of the reflexes decreases, thereby reducing the signal-to-noise ratio.

21.1.3 The Phase Problem

The X-rays diffracted from a crystal contain all the information required to deduce the three-dimensional arrangement of the atoms of the macromolecule in the crystal. This information is contained in the three parameters required to describe an electromagnetic wave: wavelength, amplitude, and phase (Figure 21.12). The wavelength of the primary beam is not altered during diffraction in the crystal lattice (coherent diffraction) and, thus, is known. Detection of the diffracted beams by an area detector allows only the measurement of the amplitude by use of the relative intensity (e.g., by the intensity of the "black" spots). A direct determination of the phase is not possible, thus an important part of the information about the arrangement of the individual atoms in a crystal is lost. Moreover, because the X-rays are diffracted at the electron hull of the atoms, X-ray crystallography does not result in exact coordinates of the nuclei/center of the atoms, but rather the three-dimensional distribution of the electrons of the atoms, the so-called electron density.

The electron density ρ at any individual point (x,y,z) in a crystal may be calculated using the structure factor amplitude $F(h,k,l)$, which is proportional to the square root of the measured intensity I for the individual reflex (h,k,l), the respective phase $\alpha(h,k,l)$, and the volume V of the unit cell:

$$\rho(x,y,z) = \frac{1}{V}\sum F(h,k,l) \times e^{i\alpha(h,k,l)} \times e^{2\pi i(hx,ky,lz)} \tag{21.3}$$

To solve this formula the values of both the amplitude and the phase for every reflex are required, in order to determine the structure of the macromolecule. The different strategies used to obtain the phase values, which cannot be measured experimentally, will be discussed in the following.

Single Isomorphous Replacement (SIR) and Multiple Isomorphous Replacement (MIR) The method of isomorphous replacement was the historical breakthrough in solving the phase problem in protein crystallography and is still an important method to determine *de novo* structures. To do so protein crystals are soaked in salt solutions or other compounds that contain an atom or molecule with a high electron density. These heavy metal ions or molecules containing a heavy atom diffuse through the aqueous channels into the crystal lattice and interact with the protein surface. Depending on the chemical properties of the metal ion or the

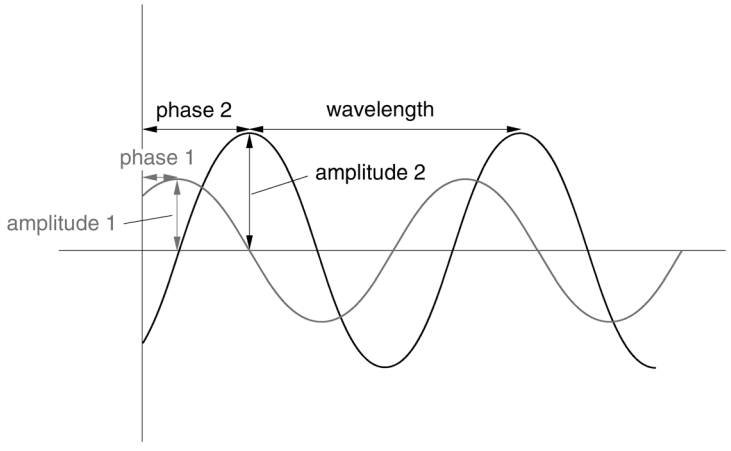


Figure 21.12 Schematic drawing of two diffracted X-ray beams having the same wavelength, but different phases and amplitudes.

metal compound they form covalent bonds with the protein, for example, Hg^{2+}-ions with the sulfhydryl group of cysteines, or the coordination of compounds like $[PtCl_4]^{2-}$ with histidine side chains. Due to their increased number of electrons, heavy atoms diffract X-rays more strongly than the atoms of the standard amino acids. Only one mercury ion bound per protein molecule already changes the relative intensities on a scattering image significantly. To use the data to solve the phase problem it is important that the heavy atom does not interfere with the protein structure and the packing of the protein molecules within the crystal, which means that the heavy atom modified crystal derivatives have to be isomorphic with the native crystal, which is devoid of any heavy atom. Thus the reflexes of the isomorph derivative crystal are at identical positions (within errors) in the respective diffraction images and only the relative intensities differ. Non-isomorphy reveals itself in changes within the cell constants or of the crystal symmetry and is a relatively frequent but undesirable phenomenon.

The differences in the intensity of the corresponding reflexes between the native and the isomorphous derivative crystal allow the calculation of the position of the heavy atoms in the unit cell of the crystal. This is possible by applying the Patterson function, which enables us to determine a crystal structure, irrespective of any *a priori* phase information, just by the use of the measured intensities. The caveat is that this method only works for simple structures with a few atoms in the unit cell. The Patterson function itself represents a set of vectors, and their norm and directions represent the distance and orientation between the atoms in the unit cell. Therefore all vectors have their starting point in the same point of origin.

For protein crystals with many atoms in the unit cell, the Patterson function cannot be applied for all atoms, but to a reduced set of atoms (the heavy atoms) using the differences between the structure factor amplitudes in the native and derivative crystal. Thus one obtains only the vectors that represent the distances and the relative orientation between the heavy atoms. Due to the crystal symmetry these vectors are correlated to each other, allowing us to derive the start and endpoints within the unit cell. In a first approximation the position of the heavy atoms within the unit cell is determined.

The phase information derived from the heavy atom positions enables us to deduce a rough estimation of the phases for the protein structure, which is called SIR (single isomorphous replacement).

The obtained SIR-phases and the measured structure factor amplitudes of the protein crystal are subsequently used to calculate an electron density map for the macromolecule. The error within the phases is still quite substantial if only one heavy atom derivative is used. The resulting errors in the electron density lead to a false or very incomplete picture of the distribution of the atoms in the unit cell. Often these SIR-electron density maps are not interpretable, meaning they lack continuous electron densities, which, for example, reveal the characteristic of an α-helix. To circumvent this problem additional data from crystals soaked with other derivatives are required. Hence, many different metal compounds are tested to find those in which other binding sites on the protein surface are used. The combination of the different phase information is therefore referred to as MIR (multiple isomorphous replacement). The advantage here is that every independent derivative reduces the error in the phase estimation, resulting in better, more interpretable electron density maps (Figure 20.13).

The error in phase determination may be further decreased by using the method of *solvent flattening*. Protein crystals contain huge channels between the protein molecules (mostly 45–70% of the total volume) which are filled with solvent. Since these water molecules do not form any ordered structure within the crystal, a significant drop in the values for electron density between the regions occupied by protein and the solvent-filled channels can be observed. Erroneous SIR- or MIR-electron density maps often exhibit small regions with increased density values within the solvent regions. These are removed using solvent flattening, by setting a constant low value for the electron density within the whole solvent channels. Subsequently, using this modified electron density distribution, the structure factor amplitudes and phases are recalculated and an improved electron density distribution is calculated by combining the new phases and the experimental structure factor amplitudes.

An alternative approach to the tedious search for multiple independent derivatives is possible if the asymmetric unit contains more than one identical protein molecule. This is often the case for oligomeric proteins, but may also occur for monomeric proteins. Based on the justifiable assumption that the structure of the monomers is identical or very similar within the asymmetric unit, the average electron density for the monomers within the asymmetric unit may be

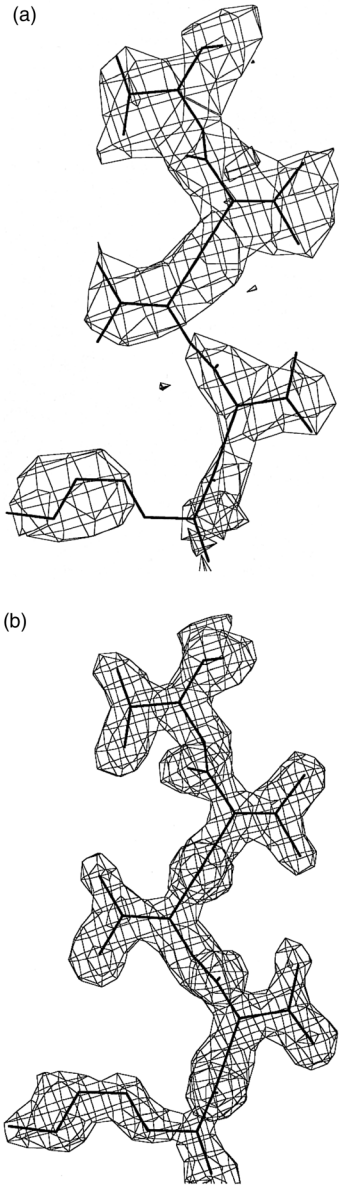


Figure 21.13 A section of the atomic model of the enzyme tRNA-guanine-transglycosylase and the respective electron density calculated in (a) with the experimental heavy atom (MIR-) phases at a resolution of 3,0 Å and in (b) with the model phases from a refined final structure model at a resolution of 1.85 Å.

calculated. To do so the position and surface have to be identified and the non-crystallographic symmetry (NCS) between the monomers has to be determined. Using these non-crystallographic symmetry operators the electron densities of the individual monomers are superposed and averaged. The higher the number of molecules within the asymmetric unit, the better the resulting averaged electron density maps.

Solvent flattening as well as NCS averaging belong to methods of density modification, since they modify existing experimental electron densities, in order to reduce the error of the phases and to improve the electron densities.

Multiple Anomalous Dispersion (MAD) Another method for the determination of protein crystal structures *de novo* is based on the anomalous scattering of X-rays. According to Friedel's law, the intensities of two reflexes, coupled by inversion symmetry $I(h,k,l)$ and $I(-h,-k,-l)$, the so-called Friedel pair, are identical in crystals containing only light atoms. The presence of heavy atoms within a crystal lattice could lead to a significant difference of the intensities of Friedel pairs. This phenomenon is described as anomalous dispersion. The difference between the intensities of the Friedel pairs is dependent on the type of atom responsible for the anomalous dispersion and on the wavelength of the X-rays. The largest effect occurs if the wavelength of the X-ray beam is close to the X-ray absorption maximum (e.g., K- or L-edge) of the metal ion, as inner electrons of the heavy atom contribute to the diffraction and to anomalous dispersion by shifting the phases.

Already, the sulfur atoms present in the amino acids cysteine and methionine of proteins cause a small anomalous effect, which may be sufficient for structure determination. Metal ions either bound as a cofactor by the protein or bound to the protein for MIR-dependent phase determination have a higher anomalous effect. From the observed differences in the intensities of the Friedel pairs the phase information can be derived.

Admittedly, the differences in intensity for most heavy atoms are quite small when using a standard rotating anode (laboratory source), which together with the standard errors in measurement results in phase information with relatively huge errors. Nevertheless the anomalous phases may be combined with the SIR or MIR phases to obtain improved phases with less error. This combination of methods is known as SIRAS and MIRAS, respectively. The effect of anomalous dispersion can be increased if the wavelength of the X-rays is tuned exactly to the absorption edge of the metal ion. Since standard rotating anodes generate only a fixed wavelength (e.g., 1.54 Å for the K_α-peak of a copper anode) another source for the generation of X-rays has to be used, namely, synchrotron radiation. Synchrotron radiation may be tuned between 0.6 and 4 Å and due to its brilliance the intensities of the reflexes can be determined more accurately.

For a complete MAD experiment (multiple wavelength anomalous dispersion) data sets at three different wavelengths are gathered from the same crystal, reducing the problem of differences in cell constants. One wavelength is chosen directly at the absorption maximum for a strong anomalous signal (peak dataset), the other (inflection point) at a wavelength close (ca. 0.001 Å) beside the absorption maximum, which is used to determine the isomorphous differences to the third dataset. This third dataset, located usually ca. 0.1 Å from the absorption maximum, serves as native dataset (high or low remote dataset). The best results are obtained if all datasets are gathered from the same crystal. To do so cryo-techniques have to be used to prevent the protein crystal from damage due to the intense synchrotron radiation. Therefore the crystal is flash cooled in liquid nitrogen and kept in a stream of N_2 gas at −100 K for the duration of the experiment. The combination of anomalous and isomorphous phases from a successful MAD experiment usually leads to an interpretable electron density map. Often, the MAD method is used with selenium atoms in recombinantly-expressed proteins. The selenium is added as selenomethionine to the growth medium of methionine-auxotrophic *Escherichia coli* cells instead of methionine, leading to the incorporation of selenomethionine instead of methionine in the overexpressed recombinant protein. This method easily replaces the long search for suitable heavy atom derivatives by soaking experiments since the selenium is already incorporated in the purified protein. This can reduce the time required from initial successful crystallization experiments to a crystal structure down to a few weeks. Thus, unsurprisingly, since the mid-1990s most *de novo*-structure determinations have been performed using the selenomethionine MAD method. The limitations of this method lie in the strength of the

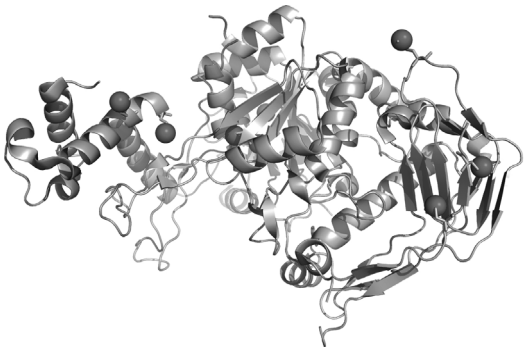


Figure 21.14 Positions of selenium atoms in the structure of the enzyme UDP-sugar-pyrophosphorylase. The methionine residues of the recombinant protein carry a selenium instead of the usual sulfur, which allows determination of the phases by SAD.

anomalous signal, which depends on the number of methionines present as well as the molecular weight of the proteins present in the asymmetric unit.

The sulfur atoms in cysteine- and methionine-side chains also reveal an anomalous dispersion, but the intensity differences of the Friedel pairs are even weaker than for selenium. Nevertheless, more and more structures are being deposited with the phase problem solved using sulfur SAD (Figure 21.14). Thus an exact as possible measurement of the intensities is very important for Se-SAD, which could be achieved by multiple measurements of each reflection. The number of the multiple measurements, the multiplicity of these datasets is often more than 20.

Molecular Replacement (MR) Besides methods for structure determination *de novo* the method of MR (molecular replacement) is frequently used if a three-dimensional structure model or parts thereof from the identical protein or from homologous proteins is known (Figure 21.15). The area of application of MR ranges from quite simple problems, like the determination of a novel crystal form of an already solved structure, through the analysis of mutant forms with known structure to the structure determination of proteins with only about 25% sequence identity to the sequence of known structures. Using the deposited coordinates of the known structure (the search model) backward calculation allows determination of the structure factor amplitudes (F_{calc}) and phases (α_{calc}) for the model. With the model phases and the experimental, observed, structure factor amplitudes (F_{obs}) an electron density for the new crystal structure can be obtained. In most cases the space group as well as the orientation of the new unknown crystal lattice differ. Therefore, in a first step the known structure has to be placed correctly in the new crystal cell. This six-dimensional search problem with three rotational and three translational variables is divided into two steps. In the first step the orientation of the molecule is determined using the Patterson function (Section 21.3.1). Initially, two Patterson

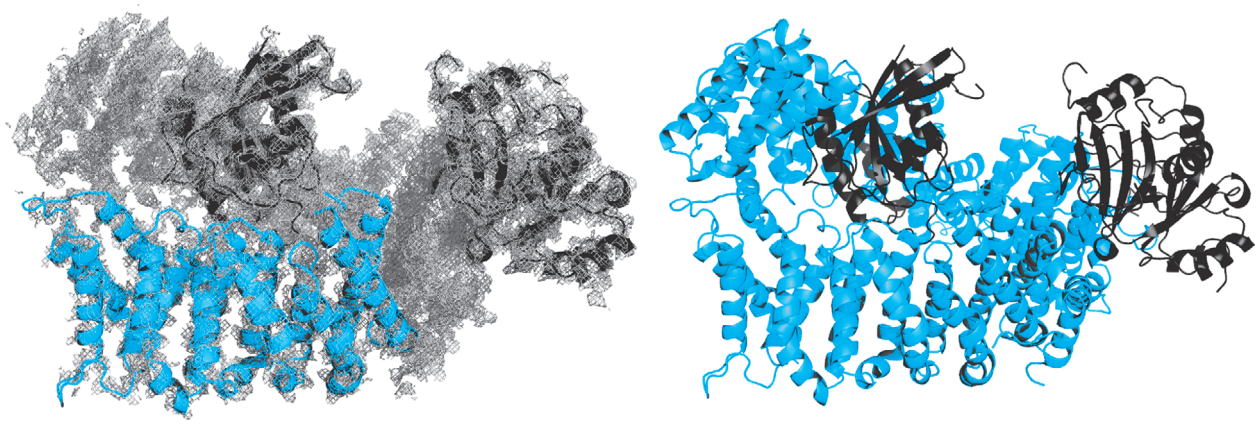


Figure 21.15 Structure determination of the CRM1-SPN1-RanGTP complex using molecular replacement. Only parts of the known CRM1 and SPN1 structures served as search models for phase determination (a). The resulting electron density (grey in (a)) allowed building of the missing parts in the structure. With increasing completeness of the structure, the quality of the electron density improves, so that a fully complete model could be formed ((b); CRM1 in blue, RanGTP and SPN1 in black).

functions are calculated, one with the observed structure factor amplitudes (F_{obs}) and the other with the calculated structure factor amplitudes (F_{calc}), based on the atomic coordinates of the search model. Subsequently, the two sets of Patterson vectors are correlated using a product function, by rotation of one set of vectors in small angular steps around all three axes (rotation function). In cases where there is a good agreement between the two sets, the rotation function obtains a high value, indicating a possible solution for the rotation. In the next step the search model is rotated into the putative solution and translated in small steps through the unit cell. After each individual step the structure factor amplitudes are calculated and correlated to the experimental ones (translation function). After placing the search model in the best orientation and position within the unit cell an electron density can be calculated and the initial structure refined.

21.1.4 Model Building and Structure Refinement

The conversion of the electron density distribution in a three-dimensional model of a protein structure is known as the interpretation of an electron density map. The increasing computer power, novel algorithms, and more profound understanding of protein structures allow initial interpretation to be performed automatically, but the final interpretation is still a manual process. Therefore the electron density map is displayed on a graphics screen and the missing residues either as polypeptide fragment or individual amino acids are placed interactively into the free electron density. The quality of the electron density map certainly depends on errors in the phases, but also on the resolution limit of the dataset. The resolution limit of the diffraction data itself and consequently also the experimental phases are limited by the properties of the crystal, especially the exactness of the crystalline arrangement. At a low resolution of about 5 Å the shape and sometimes the arrangement of α-helices as strands of electron density are barely detectable. Medium resolution (3 Å) allows us to unambiguously follow a polypeptide chain like the winding of an α-helix. Moreover the assignment of the side chains is possible at this resolution, though the amino acid sequence has to be known. At higher resolution (2 Å) more and more details become visible, such as the conformation of peptide bonds or long side chains like in arginines. Not until a resolution of 1.0 Å is reached are atoms recognizable as individual balls in the electron density.

However, few protein crystals have been obtained resulting in diffraction images with a resolution of 1.0 Å and better (0.7%), although the number has increased significantly in recent years due to developments in detectors and beam line setup. Most solved crystal structures lie within a resolution limit of 1.5 and 2.5 Å (69%), with the number of high-resolution structures between 1.0 to1.5 Å (around 9%) increasing steadily.

The initial interpretation of an electron density map is often performed at medium resolution (Figure 21.8). The resulting model usually harbors several errors, which have to be removed or at least reduced during crystallographic refinement. During the process of refinement, the difference between the experimental structure factor amplitudes (F_{obs}) and the calculated structure factor amplitudes (F_{calc}) of the actual model are minimized. Computer programs change the atom coordinates slightly, calculate new F_{calc} for the modified model, and compare them with the F_{obs}. This is an iterative process that will be repeated until the smallest possible differences between F_{obs} and F_{calc} are obtained. Refinement programs calculate the different energies, for example, of differences to the ideal geometric restraints for the individual bonds, the repulsive forces between chemically non-covalent bonds, or those describing the differences of F_{obs} and F_{calc}, and try to minimize these energy terms. At this step of refinement the resolution limit also plays an important role, since with higher resolution more data (F_{obs}) are available.

Quite frequently it is impossible to interpret a MIR or MAD electron density map completely. The erroneous experimental phases can be improved by combination with the calculated phases α_{calc} of the refined, but incomplete model. An electron density map calculated using these combined phases often reveals the missing parts, but also errors within the originating initial model. The new model is again refined against the experimental data and the model building/refinement cycle is repeated several times. With increasing completeness and refinement of the model the electron density maps are only calculated using the model phases α_{calc}, whereas mostly the differences between F_{obs} and F_{calc}, as well as the $(2F_{obs} - F_{calc})$ difference, are used

as structure factor amplitudes. The use of the $(F_{obs} - F_{calc})$ difference as structure factor amplitude brings out the missing parts of the structure model, whereas the electron density map of a $(2F_{obs} - F_{calc})$ difference Fourier synthesis reveals the complete structure model as well as the missing parts. The accordance between F_{obs} and F_{calc} is expressed in the crystallographic *R*-factor:

$$R = \frac{\sum |F_{obs}(h,k,l) - F_{calc}(h,k,l)|}{\sum F_{obs}(h,k,l)} \quad (21.4)$$

For a perfect match the *R*-factor would be zero, for an initial model it normally ranges between 0.45 and 0.5, but even for well-refined protein structures the *R*-factor lies within the range 0.15–0.20. This deviation is not due to errors in the measurement of the reflection data, but originates predominantly from the imperfect crystalline organization and the minute changes between the individual protein molecules within the crystal lattice. As a consequence of these minute differences between molecules in the crystal, the refined model represents an average of all the different conformations and orientations present in the crystal.

The flexibility of regions in the polypeptide chain that connect the secondary structure elements (α-helices and β-strands) and the resulting multiple conformations of these regions may lead to a very weak electron density map. The flexibility of the atoms within the crystal lattice are expressed by crystallographic temperature factors, the *B*-factors. They describe the misorganization as oscillation. A *B*-factor of 79 Å^2 equals an oscillation with a mean standard deviation of 1.0 Å. During the process of refinement not only the coordinates of the atoms but also the *B*-factors are varied and optimized. The refined *B*-factors, therefore, represent a measure for the flexibility of the protein structure. Values up to 20 Å^2 indicate a relatively rigid and well-defined structure. The higher the *B*-factors the poorer the atoms are defined in the electron density map.

As an additional quality control in structure refinement the method of cross validation is used. A small percentage of the diffraction data (e.g., 5% of all reflexes) is omitted from the structure refinement and used for the calculation of *R*-free, which is defined similarly to the *R*-factor. In total the quality of the crystal structure is described by multiple criteria, but lacks fixed threshold values to pinpoint rough errors in the structure.

After the refinement process and analyses of the structure the atom coordinates and the experimental data (in the form of the structure factors) are deposited in the Protein Data Bank (PDB). As soon as a publication describing the structure is available for the scientific community, the data are released and are accessible to everyone (www.pdb.org). This database also contains structures that have been determined by means of NMR or EM. A comparison of structures that have been determined using both methods, X-ray crystallography and NMR, shows that the structures are in good agreement, with the most obvious changes in the flexible regions of the polypeptide chain usually at the surface of the protein.

Besides the electron density map all values indicating a possibly wrong structure should be evaluated. These values depend on the resolution of the crystal and the other factors discussed, but as a rough estimate the following values could be used: the *R*-factor should be below 0.20, the difference between the free *R*-factor and the *R*-factor should be not more than 20% of the *R*-factor, the *B*-factors should have a mean value of 20–30 Å^2 or less and the deviation from an ideal stereochemistry should be less than 0.018 Å for the bond lengths and not more than 2.0° for the bond angles.

21.2 Small Angle X-Ray Scattering (SAXS)

In contrast to biological macromolecules in solution, which usually reveal a high degree of dynamics, macromolecules in crystal lattices are trapped in one overall conformation due to crystal packing. This rigid organization is the cause of one of the major disadvantages of crystallography. Whereas it allows us to solve structures at a high resolution the number of possible conformations is strongly reduced to a few or even only one single conformation. These might be even further reduced by the cryo-cooling used to prevent X-ray damage. Exceptions with respect to flexibility are, usually, the side chains of residues located on the surface or loops that are pointing into the solvent channels and are not involved in crystal packing contacts. They are still able to adopt multiple conformations.

Proteins or protein complexes in solution usually reveal a high degree of flexibility and/or dynamics, meaning multiple conformations are in equilibrium with each other. Interaction with binding partners or changes in the environment affect the equilibrium, and thus the amounts of the individual conformations. This total population or ensemble of conformations can be investigated in solution by means of SAXS (small angle X-ray scattering). SAXS allows us to gather information of a thermodynamic ensemble, resulting in an averaged information of all

states of the molecule in solution, such as particle size, shape, and ratio of surface to volume. Thus, SAXS increases the "conformational and structural space" of the rigid structures obtained by X-ray crystallography, enabling the determination of structural changes due, for example, to interaction with other molecules. For example, crystal studies have shown the interaction of the DNA repair enzyme Mre1 with DNA and described the interaction of a second Mre1 molecule as caused by crystal contacts. The SAXS experiments identified this interaction as important for dimer formation and essential for the DNA binding (Williams, R.S., *et al.* (2008) *Cell*, **135**, 97–109).

Overall, small angle scattering can be seen as a complementary method between X-ray crystallography and NMR spectroscopy with their structure models at atomic resolution, on one hand, and single particle EM and tomography with their resolution and size limitations (allowing investigation of only molecules of greater molecular weight than 100 kDa) on the other hand.

21.2.1 Machine Setup

Optimal results for SAXS measurements are directly correlated to a brilliant beam source with a divergence-free, parallel X-ray beam. This may be achieved either by using mirror systems that, by bending the beam, focus a divergent beam or by using collimators, for example, like an array of apertures that exclude divergent beams and allows only the parallel beam to reach the sample. Advantageously, the array of apertures is located directly before the sample in the beam, thereby also eliminating unwanted background dispersion. X-Rays pass through the sample, where they are scattered, and are measured by a detector located behind the sample (Figure 21.16). The smaller the angle that should be detected and the intensities measured, the more the detector to sample distance has to be increased to ensure unambiguous separation of the primary beam directly passing though the sample without any interference and the scattered beam (Figure 21.16). The beam travels in a high vacuum tube to prevent any undesired interaction and thus scattering with air molecules. The detector measures the angle dependent intensity of the beam scattered by the sample. As SAXS detectors, modern hybrid pixel array or CCD cameras are used, as in X-ray crystallography.

The quality of the sample is crucial for the quality of the results. On the one hand, the best concentration for the experiment has to be determined (0.1–10 mg ml^{-1}). On the other hand, the sample should contain no aggregates or misfolded proteins. This problem can be circumvented if the sample is centrifuged to precipitate large and thus heavier aggregates. Moreover, no interaction between the molecules should be present. This information may be obtained by performing experiments at multiple concentrations, which in the ideal case of no interaction

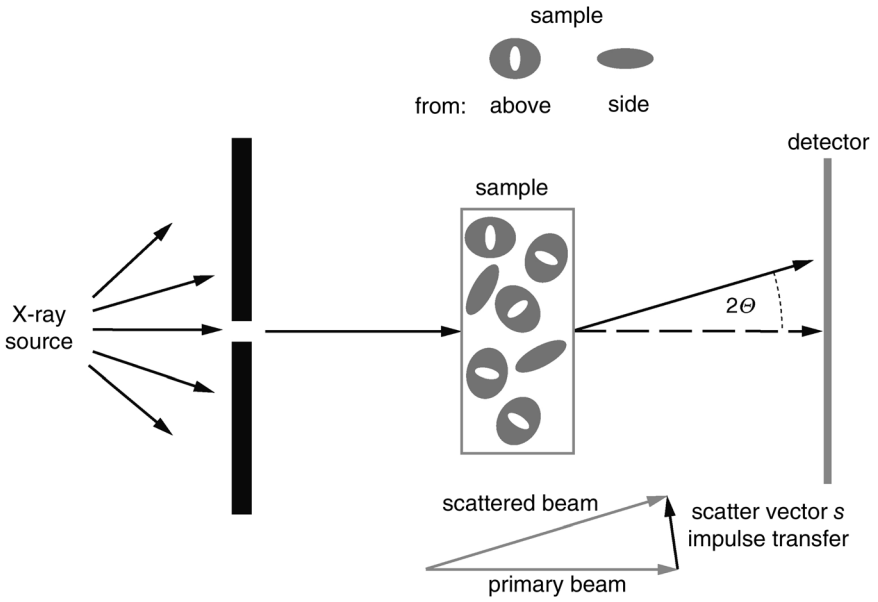


Figure 21.16 Schematic representation of a SAXS experiment. X-Rays from the source (rotating anode or synchrotron) are separated from divergent beams by collimators or mirror systems before hitting the sample. A huge percentage of the rays pass through the sample directly and are absorbed by the "beam stop." The scattered beam is measured by a detector; 2Θ denotes the diffraction angle. The distance between the primary beam and the diffracted beam on the detector is proportional to the scattering vector *s*.

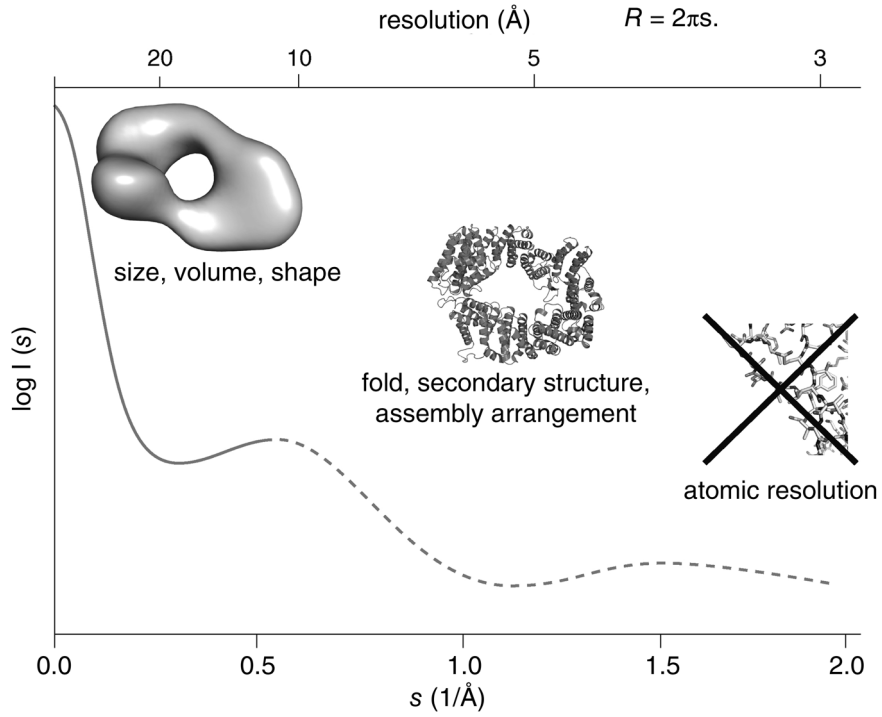


Figure 21.17 Schematic representation of the scattering curve resulting from a SAXS experiment. The intensity (I) is plotted against the scattering vector s. The dashed line indicates the regions in which scattering is hard to observe, if at all. In parallel with the scattering vector the resolution R (with $R = 2\pi/s$) is shown above the graph. The representations of structures illustrate the quality of the structural information that may be deduced at the respective scattering vector (or resolution). The models depicted to demonstrate the structural information gained at the respective resolution where obtained using two structure models of the export receptor CRM1 (chromosome region maintenance 1) (EMdb EMD-1099 and PDBid: 3GJX).

should result in identical values. For SAXS experiments there are almost no size restrictions for protein/macromolecules or protein complexes ranging from small molecules in the low kDa range to MDa sizes as found for virus particles or ribosomes. The influence of the surrounding medium on the observed scattering distribution is determined by measurement of blank values that contain the identical buffer the sample is dissolved in. Subtraction of these values from the values of the sample measurement results in a curve caused only by the scattering of the macromolecule in solution. For subsequent calculations and meaningful evaluation of the data a molecular weight standard is measured as well, usually a globular protein of known structure (e.g., bovine serum albumin), and used for calibration.

An established representation of the obtained intensity distribution is the logarithmic plot of the intensity versus the scattering vector s (Figure 21.17). Such scattering curves have some common features. At small s-values (0–0.4 $Å^{-1}$, small angles, resolution range 50–15 Å) a steep decent of the curve is observed. This characteristic of the curve is mainly determined by the three-dimensional shape of the sample. The region is used to determine size, shape, and volume of the sample and reveals the greatest differences between samples. The differences in the neighboring range (15–5 Å) are not that significant between curves of different samples; at higher resolution range (below 4 Å) the curves are almost identical. Due to this fact only limited information with respect to the organization of secondary structure elements may be made at an atomic level.

An improvement is the method of wide angle X-ray scattering (WAXS), which allows us to detect signals that are required for interpretation in the resolution range 4–3 Å. With a high resolution structure in hand, this enables us to detect structural differences in the center of the molecule investigated, such as, for example, due to ligand or protein binding.

21.2.2 Theory

SAXS is based on elastic scattering of monochromatic X-rays ($\lambda = 1$–2 Å) on the surface of molecules, which is detected at small angles (0.1–10°) with respect to the beam axis. Within this range of the scattering angle conclusions about the shape (radius of gyration (R_g), hydration volume (V_h), maximal diameter (D_{max}), and molecular mass (M_w)) of the molecule may be drawn.

Under optimal conditions – meaning best dilution, prevention of aggregate formation, no buffer mismatch between sample and blank (buffer only) – the scattering curve displays the

sum of all scattering profiles (with intensity I) of all molecules contributing to the scattering. Because the macromolecules are free in solution, they can obtain any orientation with respect to the X-ray beam. This is also the reason why, in contrast to X-ray crystallography, the maximal reachable resolution is significantly lower. For representation purposes the intensity of the scattering (I) is plotted as a function of the scattering vector (s). The scattering vector s is defined by:

$$s = 4\pi(\sin\theta)/\lambda \tag{21.5}$$

where θ is the scattering angle in relation to the primary beam and λ is the wavelength of the X-ray beam used. Thus, the value for the scattering vector allows us to gain information about the properties of the macromolecule independent of the wavelength. To correlate the value of the scattering vector to only the macromolecule, the result and therewith the scattering has to be corrected by the amount of solvent in which the macromolecule has been measured.

The intensity distribution of a SAXS experiment can be described as follows:

$$I(s) = P(s) \times S(s) \tag{21.6}$$

where $I(s)$ is the intensity according the respective scattering vector, which in turn is composed of $S(s)$, the structure factor describing the organization between the macromolecules and $P(s)$, the form factor describing the shape of the macromolecule. Common to all form factors is the property that they may easily be converted from the real into the reciprocal space by a Fourier transform of the spatial intensity distribution.

With the help of the resulting scattering curve it is possible to extrapolate the intensity distribution of an individual molecule for a monodisperse solution of macromolecules that are devoid of any attractive or repulsive forces. Only under these conditions is the structure factor $S(s)$ equal to 1 and, thereby, the form factor $P(s)$ alone determines the intensity. In this case the measured overall intensity is proportional to the intensity of a single molecule, but with the shape averaged over all conformations of the macromolecule found in the sample.

For small s-values Guinier's law may be applied, which in a first approximation correlates the intensity to the radius of gyration (R_g) of the macromolecule. The R_g describes the mass distribution of a macromolecule in relation to its center of mass and equals the square root of the averaged squared distance of the individual atoms to the center of mass of the total macromolecule. For a given protein an increase in R_g reflects an opening of the molecule, whereas a decrease of R_g indicates a reduction in size. The measured SAXS data may be presented in different plots, varying in the kind of axis definition and allowing different graphical presentations and interpretation of the experimental data.

For the Guinier plot (Figure 21.18a) the logarithmic representation of the intensity [$\log(I(s))$] versus the square of the respective scattering vector (s^2) is used. In the ideal case, with no aggregates (attractive forces) present in the solution, this plot results in a straight line with its gradient permitting deduction of the R_g (for values of $sR_g < 1.3$). From such plots, aggregation can be detected as a bad fit to a straight line. The extrapolated intensity for $s = 0$ provides

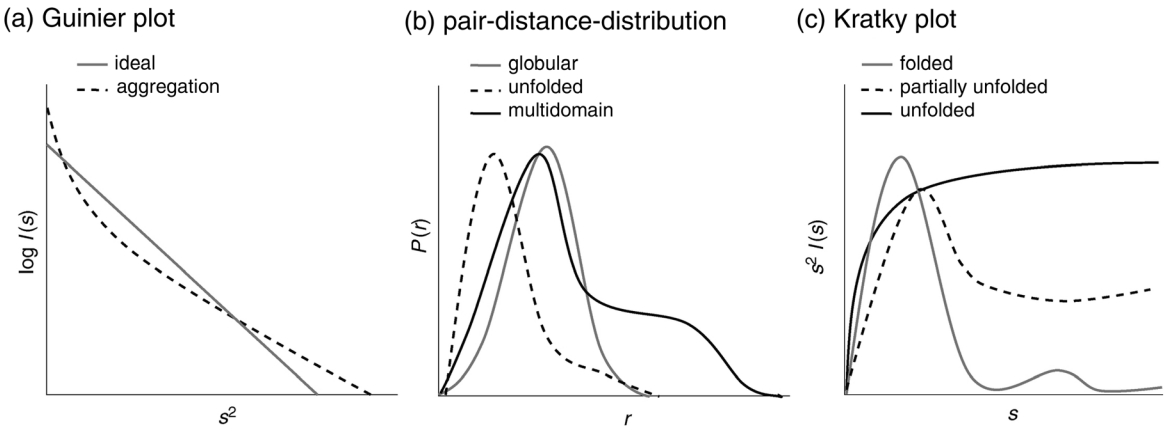


Figure 21.18 Different ways of presenting SAXS data and their interpretation. (a) Guinier plot, (b) pair distance-distribution function, and (c) Kratky plot. See text for details.

information on the molar mass (M_w) of the sample and may be used for mass determination if a calibration with a reference has been performed.

An elegant way to gain insight into the overall structure or folding state of a molecule and of presenting the conformational changes of a molecule according to external influences or upon ligand binding is possible using the pair-distance distribution ($P(r)$ function) (Figure 21.18b). The $P(r)$ function represents the distances of all possible combinations of "dummy atoms" (see below for explanation) within the object of interest. The number of pairs of dummy atoms with a defined distance is plotted in relation to the distance (r). The spatial rearrangement of a small number of residues or atoms within a molecule is already sufficient to cause a significant change in the $P(r)$ function. For globular proteins the pair distribution function $P(r)$ plotted against the distances of the pairs (r) would result in an almost perfect bell shape of a Gaussian distribution. The largest value for r would represent D_{max}. The plot for an unfolded protein would reveal a shift of the maximum towards a smaller r as well as a more gradual descent towards bigger r values. This is caused by the increased number of interactions between neighboring molecules that usually are in close proximity. Multidomain proteins might also exhibit a shift of the maximum towards smaller r values, since the interactions within the single domains form most of the contacts. In addition, a characteristic of multidomain proteins is intermediate plateaus and local maxima, which indicate the individual separated domains. Finally, changes within the structure of a molecule due to ligand binding or changes in the buffer condition can also be detected in the shape of the curve of the pair distribution function.

For proteins with a high degree of folds Prorod's law may be employed to determine the hydration volume (Porod volume) of a molecule using the primary SAXS data. This approach is based on the idea that at a reasonably high resolution the contributions to the scattering curve depend only on the transition region between solvent and protein in which the solvent is organized in a sort of ordered fashion around the protein. Independent of the shape of the molecule, the size of the surface of all proteins in the sample can be determined.

More detailed information on the folding state of a molecule may be deduced from the Kratky plot (Figure 21.18c). Here, the product of the squared scattering vector and the intensity ($s^2I(s)$) is plotted against the scattering vector s. This representation enables a simplified interpretation of the folding state of the molecule. A well-folded molecule reveals the devolution of a closed bell shaped curve with side maxima and approaches zero for high s-values. Unfolded proteins exhibit the shape of a hyperbolic curve, which asymptotically reaches a maximal value. Partly folded molecules exhibit intermediate curve behaviour.

21.2.3 Data Analysis

***De Novo* Structure Determination** As it is limited by a resolution of about 10 Å, SAXS so far enables the *de novo* or *ab initio* structure determination only with a more or less detailed representation of the molecule surface (Figure 21.18). Initial calculations of the surface were based on the fit of spherical models to the obtained data by use of additional parameters deduced from the data. Therefore, the resolution range 15–30 Å (scattering vector to 0.4 $Å^{-1}$) is used (Figure 21.19) and by statistic averaging the most plausible structure is determined.

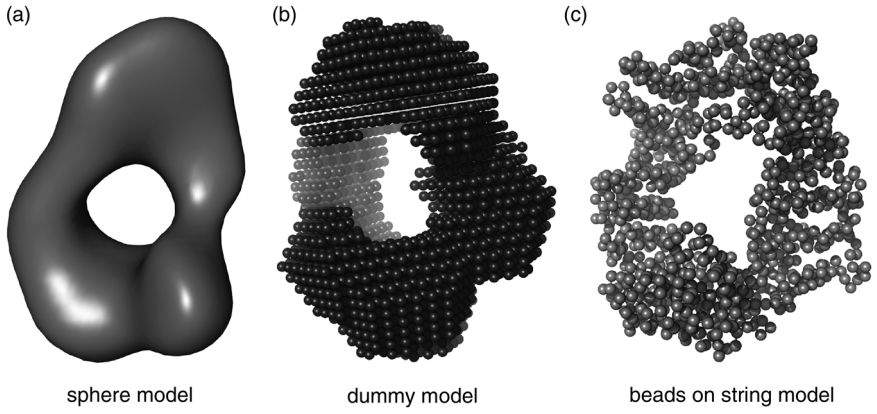


Figure 21.19 *Ab initio* structure determination. Schematic representation of results using different methods: (a) sphere model, (b) dummy model, and (c) beads on a string model. The models where compiled utilizing the EM and crystal structures of CRM1 (EMdDB EMD-1099 and PDBid: 3GJX) or calculated using measurements of CRM1 with a fast mode of the program DAMMIF.

More recent methods use so-called dummy residues (DR) as place holders that represent, depending on the size, one or more residues of the molecule and therefore indirectly the resolution obtained for the final model. The DRs are placed into the molecule under the restraint of the data curve, so that the sum of all DRs present in the model represents the overall shape of the molecule. Adaptation to the obtained SAXS data is achieved by using the method of "simulated annealing," which allows in a relatively short time the generation of a model that best fits the obtained data. In this heuristic approach the positions of an ensemble of DRs is optimized (cool down phase) until a (local) minimum is reached. Subsequently, the model is enabled to occupy similar but less optimal structures (heating phase), to allow the DRs to reorganize before it is cooled down again. This procedure prevents the DR model becoming stuck in a local optimization minimum and it may reorganize finally into a plausible structure at the minimum. The resulting model nicely represents the molecule's surface, but it is a model without any structural details at atomic resolution.

An advancement is to represent the DRs as beads on a string to mimic the actual polypeptide chain (dummy protein method). For this method, curve information up to 5 Å resolution can be used. A requirement is knowledge of the number of amino acid residues in the molecule.

A quite different approach to determine structures using SAXS data is based on the use of high-resolution structure models, which are adapted to their respective SAXS curve by "rigid body modeling" (Figure 21.20). Here, the structures are used as rigid elements with a defined shape and size that are fitted into the model obtained from SAXS. Having determined the best-fitting orientation, the details may be optimized by molecular dynamics (MD) simulation.

An interesting advantage of the SAXS method is the possibility to take the surface models generated from the scattering curves and fit high resolution structures of proteins or domains into them. Predominantly, surface models of multidomain proteins with a high degree of flexibility of the domains with respect to each other or macromolecular complexes of proteins are of interest. Especially for the latter, SAXS measurements of the individual components or subcomplexes aid, by a subtractive modeling approach, localization of the region where a protein is positioned within the complex – this region can subsequently be filled with a high-resolution structure. With this elegant approach the high-resolution structures of proteins or domains may be combined with the low-resolution information of complexes or multidomain proteins.

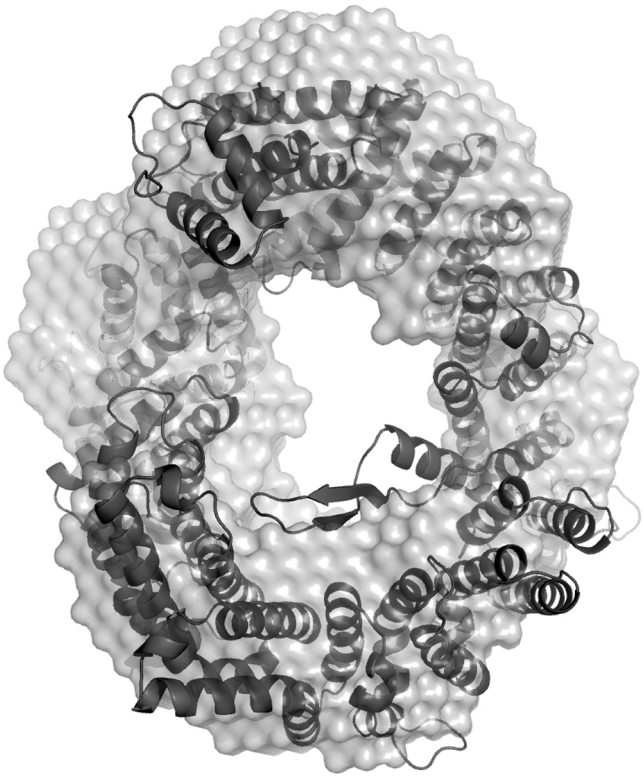


Figure 21.20 *De novo* structure determination and fitting of known structures from the Protein Data Bank (PDB).

Outlook: Method Developments More recently developed methods enable the expansion of the analysis in regions such as missing fragments, the determination of differences in a population of oligomers, and the integration of flexible regions like loops into structures. In high-resolution structures obtained by, for example, X-ray crystallography, loop regions or even whole domains might be missing due to the flexibility of these regions compared to the remainder of the protein. In contrast, the data obtained by SAXS still contain this information, allowing us to deduce and incorporate the most probable arrangement of the missing part or residues (Petoukhov, M.V. *et al.* (2002) *Biophys. J.*, **83,** 3113–3125).

A substantial disadvantage of SAXS is caused by the sample, as the SAXS data derive from an ensemble that is structurally often heterogeneous and is usually is made up of some trillion molecules in a typical sample volume. Recent developments in computing power and *in silico* structure modeling, or calculations like MD simulation and 3D modeling, have opened up novel ways to obtain detailed information about proteins. Another interesting step in this direction is the ensemble optimization method (EOM, Bernado, P., *et al.* (2007) *J. Am. Chem. Soc.*, **129**, 5656–5664), which sets up a model based on known structures or parts of molecules and subsequently models the missing links. Many possible conformations may be generated and then analyzed for the best fit to the obtained SAXS data. In the minimal ensemble search (MES) method (Pelikan, M., Hura, G.L., and Hammel, M. (2009) *General Physiol. Biophys.*, **28** (2), 174–189) this goal is approached by using MD simulations at high temperature to obtain different possible models and subsequently generic algorithms in order to reduce the number of members that fit in the ensemble of structures so as to fulfil the restraints of the SAXS data.

21.3 X-Ray Free Electron LASER (XFEL)

A highly interesting upcoming method is based on the use of a free electron LASER (FEL) with the ultimate goal of determining the three-dimensional structure of single proteins or complexes in solution at high resolution. An intermediate setup, ranging between the crystals used in X-ray crystallography and single molecules in solution, has been achieved already by using microcrystals. Even small crystals have the capability to increase the diffraction pattern significantly due to their repetitive arrangement of the molecules. Such nanocrystals were sufficient to determine the structures of lysozyme and the cysteine protease cathepsin B (Boutet, S. *et al.* (2012) High-resolution protein structure determination by serial femtosecond crystallography. *Science*, **337**, 362–364., Redecke, L. *et al.* (2013) Natively inhibited *Trypanosoma brucei* cathepsin B structure determined by using an X-ray laser. *Science*, **339**, 227–230).

The principle of the FEL is similar to that of a classical LASER (light amplification by stimulated emission of radiation) in terms of obtaining a coherent electromagnetic beam. Electrons are excited and release – depending on the experimental setup – rays of a defined wavelength. In contrast to a classical LASER, with the electrons bound to a medium, the FEL uses a relativistic (close to the speed of light) moving beam of electrons that traverses a free electromagnetic field. The emitted wavelength of a FEL is variable in a wide range. The FLASH (free electron laser in Hamburg) at DESY emits light in the UV-range and soft X-rays (6–30 nm). X-Ray beams of 0.15 nm are available to date (2014) at multiple sites (http://sbfel3.ucsb.edu/www/vl_fel.html) like the LCLS (Linac Coherent Light Source) in Stanford (USA) while SACLA (SPring-8 Angstrom Compact Free Electron Laser) in Japan started operations at 0.063 nm Ishikawa, T. *et al.* (2012) A compact X-ray free-electron laser emitting in the sub-ångström region. *Nature Photon*, **6**, 540–544.). In 2014 the European XFEL is scheduled to commence operations with a wavelength of 0.005 nm.

21.3.1 Machine Setup and Theory

The general setup of the free electron LASER requires, principally, only a few components (Figure 21.21). The electron beam has to be generated in an accelerator. In a second step the electrons are further accelerated to a relativistic speed. Subsequently, the electrons, as bunches or pulses, pass through an undulator. Within the undulator the alternating orientation of the poles of magnets leads to a transversal reorientation of the electrons at high frequency, resulting in the generation of high energy synchrotron radiation. The relativistic speed of the electrons

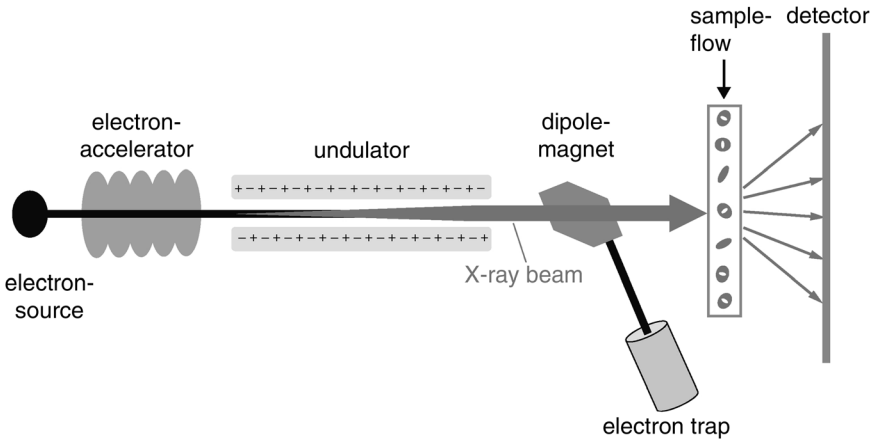


Figure 21.21 Schematic drawing of a free electron laser (FEL) setup.

causes the radiation to be almost completely oriented parallel to the electron beam. Only in the case of a perfect match of the speed of the electron bunch within the magnets of the undulator and the emitted radiation may a constructive interference of the emitted light waves take place, resulting in its desired amplification. The pulses of electrons and the specific properties of the undulator result in X-rays of a pulse duration of a few femtoseconds.

After the undulator a dipole magnet separates the electron beam from the X-ray beam before the X-ray beam hits the sample. The resulting diffraction image is recorded by novel fast readout detectors and streamed into the data processing pipeline (Strüder, L. *et al*. (2010) Large-format, high-speed, X-ray pnCCDs combined with electron and ion imaging spectrometers in a multipurpose chamber for experiments at 4th generation light sources. *Nucl. Instrum. Methods*, **614**(3), 483–496; Denes, P. *et al*. (2009). A fast, direct x-ray detection charge coupled device. *Rev. Sci. Instrum*. **80**(8), 083302.). As in X-ray crystallography and SAXS experiments, analysis of the data is dependent on computers and specialized software.

Samples Similar to the SAXS experiments the ultimate goal is the use of samples in solution, but so far nanocrystals have been used. Nanocrystals can significantly increase the diffraction pattern due to their repetitive arrangement of molecules. Due to the high energy dose a single pulse disintegrates the sample and melts it to plasma. Therefore, only one image may be taken before the sample is destroyed. To obtain a full dataset a flow of a highly diluted solution of particles (to date, nanocrystals) is directed through the beam.

Detection and Analysis With respect to sensitivity and readout time novel detectors had to be developed. Their sensitivity is suitable for the detection of single photons and they can eliminate undesired stray and plasma radiation, resulting in brilliant diffraction images.

The method of nanocrystallography is based on the same principles as X-ray crystallography and solves the phase problem in a similar manner. Datasets are gathered by means of serial femtosecond crystallography (SFX). In contrast to X-ray crystallography where a dataset is usually measured from one crystal, datasets obtained by FEL require thousands of nanocrystals since only one image may be taken per crystal. This was called a "diffraction-before-destruction" approach (Redecke, L. *et al*. (2013) *Science*, **339**, 227–230). Because the nanocrystals are in different orientations during the illumination the diffraction images represent all possible orientations of the diffraction pattern. To obtain a dataset a huge number of images has to be taken. The position and intensity of the spots on the individual images are gathered at the highest accuracy possible, indexed, sorted with respect to orientation of the molecule in the beam, and scaled to form a dataset. Subsequent analysis is similar to the approaches described above.

Acknowledgement

We wish to thank Dr. Thomas Monecke and Dr. Piotr Neumann for critically reading this chapter.

Further Reading

X-Ray Crystallography

Carter, C.W. and Sweet, R.M. (eds) (2003) *Macromolecular Crystallography, Part C*, Methods in Enzymology, vol. 368, Elsevier Academic Press, San Diego.

Carter, C.W. and Sweet, R.M. (eds) (2003) *Macromolecular Crystallography, Part D*, Methods in Enzymology, vol. 374, Elsevier Academic Press, San Diego.

Messerschmidt, A. (2006) *X-Ray Crystallography of Biomacromolecules – A Practical Guide*, Wiley-VCH Verlag GmbH, Weinheim.

McPherson, A. (2002) *Introduction to Macromolecular Crystallography*, John Wiley & Sons, Inc., Hoboken.

Rhodes, G. (2000) *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*, 2nd edn, Academic Press, New York.

Rupp, B. (2010) *Biomolecular Crystallography, Principles, Practice and Applications to Structural Biology*, Garland Science.

SAXS

Gomez, M., Nogales, A., Garcia-Gutierrez, M.C., and Ezquerra, T.A. (eds.) (2009) *Applications of Synchrotron Light to Scattering and Diffraction in Materials and Life Sciences*, Lecture Notes in Physics, vol. 776, Springer.

Putnam, C.D., Hammel, M., Hura, G.L., and Tainer, J.A. (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Quarterly Rev. Biophys.*, **40** (3), 191–285.

Svergun, D.I. and Koch, M.H.J. (2003) Small-angle scattering studies of biological macromolecules in solution. *Rep. Prog. Phys.*, **66**, 1735–1782.

Rambo, R.P. and Tainer, J.A. (2010) Bridging the solution divide: comprehensive structural analyses of dynamic RNA, DNA, and protein assemblies by small-angle X-ray scattering. *Curr. Opin. Struct. Biol.*, **20**, 128–137.

FEL

Ravasio, A. *et al.* (2009) Single-shot diffractive imaging with a table-top femtosecond soft X-ray laser-harmonics source. *Phys. Rev. Lett.*, **103**, 028104.

Saldin, E.L., Schneidmiller, E.V., and Yurkov, M.V. (2000) *The Physics of Free Electron Lasers*, Advanced Texts in Physics, Springer Verlag, Berlin, Heidelberg.

Barty, A., Küpper, J., and Chapman, H.N. (2013) Molecular imaging using X-ray free-electron lasers. *Annu. Rev. Phys. Chem.*, **64**, 415–435.

Part III

Peptides, Carbohydrates, and Lipids

Analytics of Synthetic Peptides

22

Annette Beck-Sickinger and Jan Stichel

Leipzig University, Institute of Biochemistry, Brüderstraße 34, 4103 Leipzig, Germany

Peptide synthesis is of great importance for the evaluation of bioactive peptides like hormones and neurotransmitters as well as in protein research. Designed and subsequently synthesized protein segments serve as reference compounds for the identification of the primary structure and the evaluation of the secondary structure of proteins and protein segments. Additionally, synthetic peptides are also used for the characterization of epitopes of viral and bacterial surface proteins ("synthetic vaccines") and for the generation of antibodies against specific protein domains. As fluorescence techniques have evolved rapidly in recent years, selectively labeled fluorescent peptides are often used to investigate peptide–protein and protein–protein interactions. For simple and efficient synthesis of the peptides, a solid-phase supported synthesis strategy is used, particularly if only limited amounts (<10 mg) with medium purity (95–98%) are needed.

22.1 Concept of Peptide Synthesis

The chemical synthesis of peptides and proteins is based on a stepwise condensation reaction of amino acid building blocks (Figure 22.1). As each amino acid carries an amino as well as a carboxyl function, these groups have to be selectively protected during the synthesis by reversibly cleavable protecting groups (Figure 22.2). To completely avoid undesired side reactions, the functional groups on the side chains of the so-called trifunctional amino acids (Lys, Arg, His, Glu, Asp, Ser, Thr, Tyr, Cys) also have to be reversibly protected.

However, for an efficient condensation reaction, the free carboxyl group must be activated to be reactive.

The most commonly used method for peptide synthesis was introduced by Bruce Merrifield in 1963. The peptide is sequentially synthesized on a polymer support from the C-terminus to

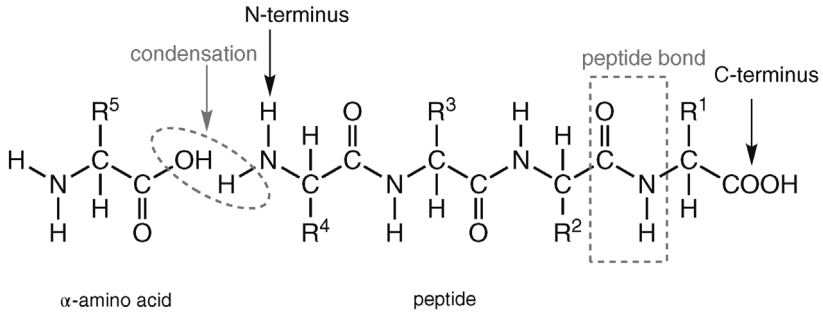


Figure 22.1 Principles of peptide synthesis. The peptide bond, which connects the amino acids, is formed by a condensation reaction.

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

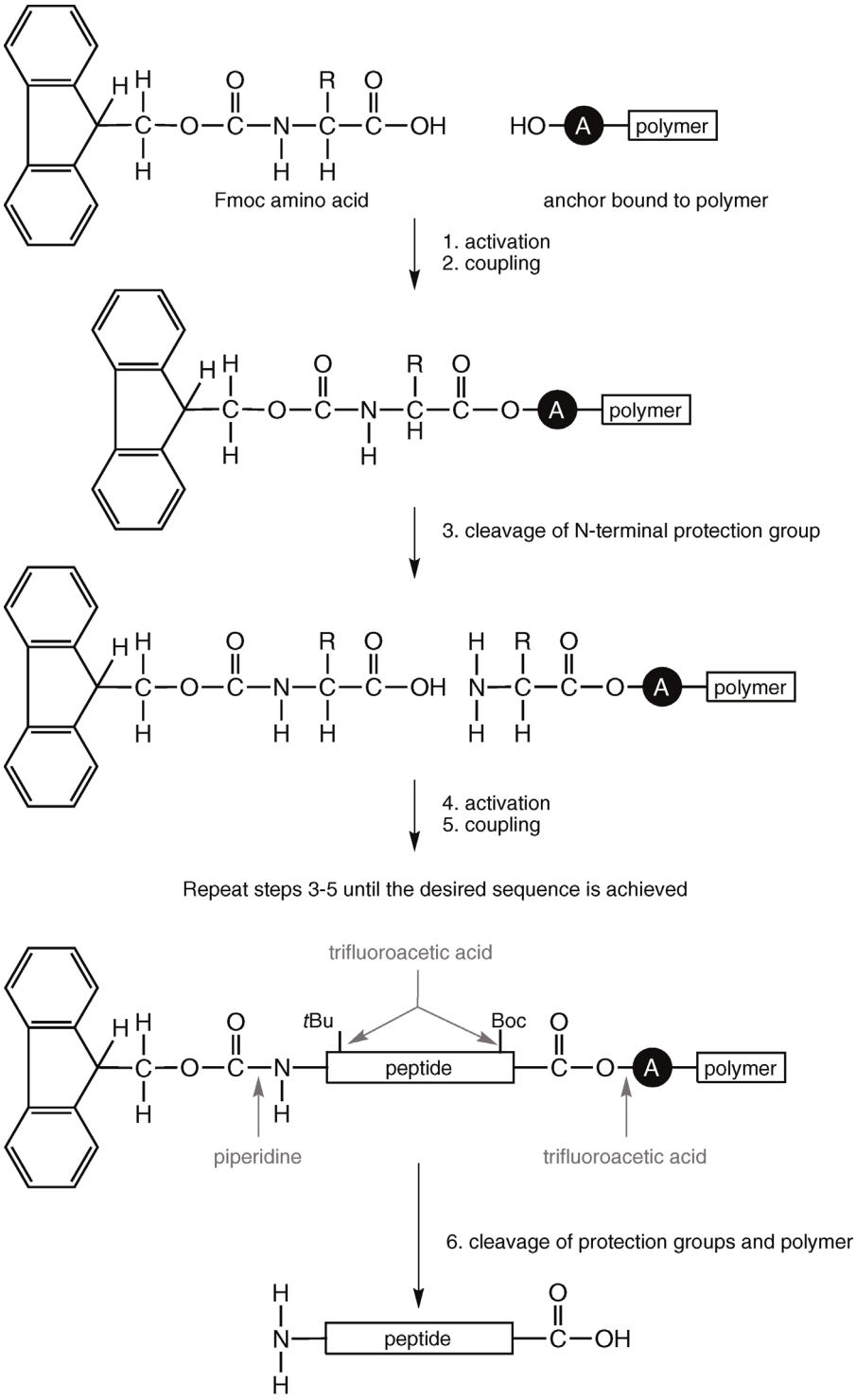


Figure 22.2 Synthesis of peptides on a solid support. The peptide is covalently bound to the resin via an anchor. The N-terminally protected amino acids are activated and coupled to the nascent peptide. The N-terminal protecting group is cleaved and the cycle is repeated until the desired length is achieved. In the last step, the side chain protecting groups are split off and the peptide is released from the resin.

the N-terminus. In the first step, the carboxyl group of the C-terminal amino acid is coupled to an anchor function on the polymer support (resin) (Figures 22.2 and 22.3). This anchor function allows the selective cleavage of the fully synthesized peptide from the solid phase using defined conditions. The next amino acid is N-terminally protected, C-terminally activated (e.g., as ester), and coupled to the free amino group of the peptide chain. Subsequently, the protecting group of the amino acid is cleaved and followed by the coupling of the next amino acid. This cycle of activation, coupling, and deprotection is repeated until the peptide reaches the desired

length. The peptide is now cleaved from the resin (the covalent bond between C-terminal amino acid and anchor function of the resin is cleaved), which yields the peptide as acid, amide, or any other modification, depending on the anchor function. Usually, the side chain protecting groups are also cleaved in this step.

When choosing the side chain protecting groups, it needs to be considered that the N-terminal protecting group has to be cleaved several times during synthesis without removing the side chain protecting group. Hence, a graduated stability of the N-terminal versus side chain protecting groups is required. However, a certain lability of the side chain protecting groups has to be kept for cleavage in the final synthesis step without damaging the peptide. Usually, these requirements are met by the Fmoc- or Boc-strategy. Figure 22.2 shows the synthesis concept according to the Fmoc-strategy.

Using the Fmoc-strategy, the α-amino group is protected by a fluorenylmethoxycarbonyl (Fmoc) moiety, which is cleavable under basic conditions. The Fmoc group is coupled to the α-amino group by a urethane bond and can be cleaved by piperidine (β-elimination) (Figure 22.4a). The lability under basic conditions of this bond is caused by the high acidity of the C_9-H of the fluorine system and the relatively high stability of the resulting anion. To protect the side chain, acid labile protecting groups are used accordingly. The anchor, which links the peptide to the resin, is usually also cleavable under acidic conditions.

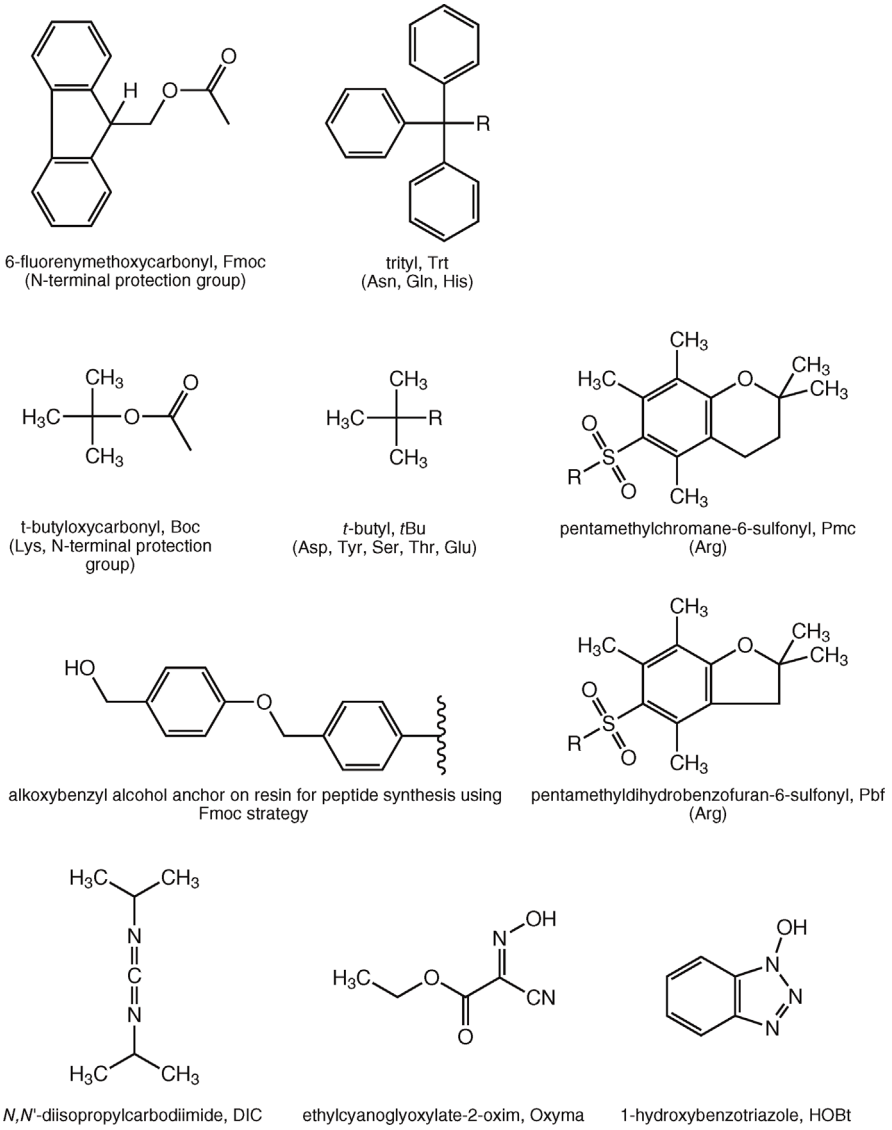


Figure 22.3 Structure and abbreviation of selected agents used in peptide synthesis: N-terminal protecting groups, side chain protecting groups, anchors, and coupling reagents.

Subsequent to cleavage of the Fmoc group at the N-terminus of the peptide chain, the carbon atom of the carboxyl function of the following amino acid has to be activated by electron-withdrawing groups. In the past, this has usually been realized using *N,N′*-diisopropylcarbodiimide (DIC) and 1-hydroxybenzotriazole (HOBt). As anhydrous HOBt is explosive and thereby under legal limitations with respect to transportation, HOBt has increasingly been replaced by other activation reagents. The most promising compounds are ethylcyanoglyoxylate-2-oxime (Oxyma) and its derivatives.

For the coupling of the amino acid to the peptide chain, the amino acid is added in fivefold excess and activated using DIC/Oxyma. Subsequently, the resin is washed using dimethylformamide (DMF) and the amino acid is again added in fivefold excess and is activated by DIC/Oxyma (Figure 22.4b). The activation mechanism using DIC includes an O-acylisourea as intermediate product, which can then react with the amino group of an amino acid to form a peptide bond. In the presence of Oxyma, the O-acylisourea is converted into O-acyloxime, which exhibits a lower tendency to form an enol at the chiral C_α. Hence, the risk of racemization is significantly lowered. To reduce the amount of false sequences and to increase the yield, one (or more) positions may be coupled twice without cleaving the N-terminal protecting group.

The cleavage of the protecting groups and the cleavage of the peptide chain from the resin is conducted in one step using concentrated trifluoroacetic acid (TFA) and 5–20% scavenger. Scavengers are compounds that are highly reactive with cations and radicals and thereby

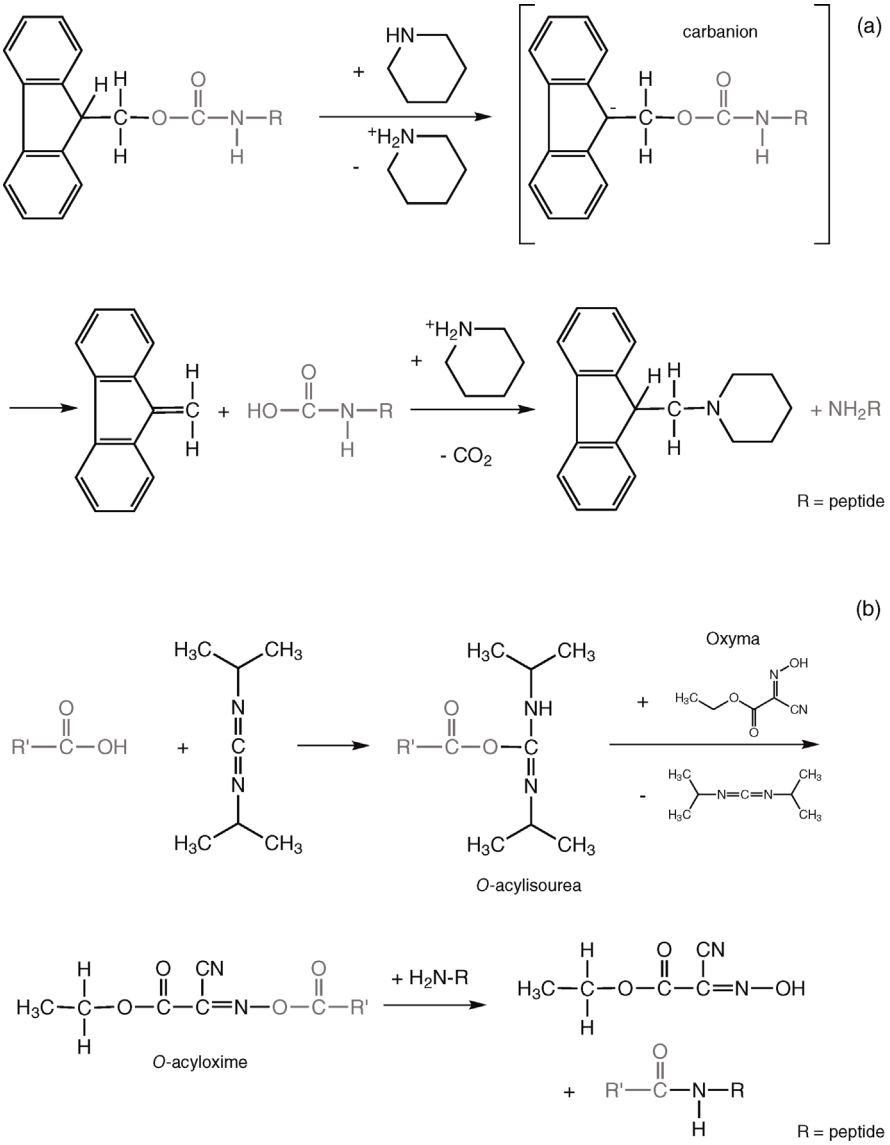


Figure 22.4 Important reaction mechanisms in peptide synthesis. (a) Cleavage of the N-terminal protecting group Fmoc by piperidine. (b) Activation of the carboxyl group by DIC and Oxyma and coupling of the amino acid (formation of a peptide bond by a condensation reaction).

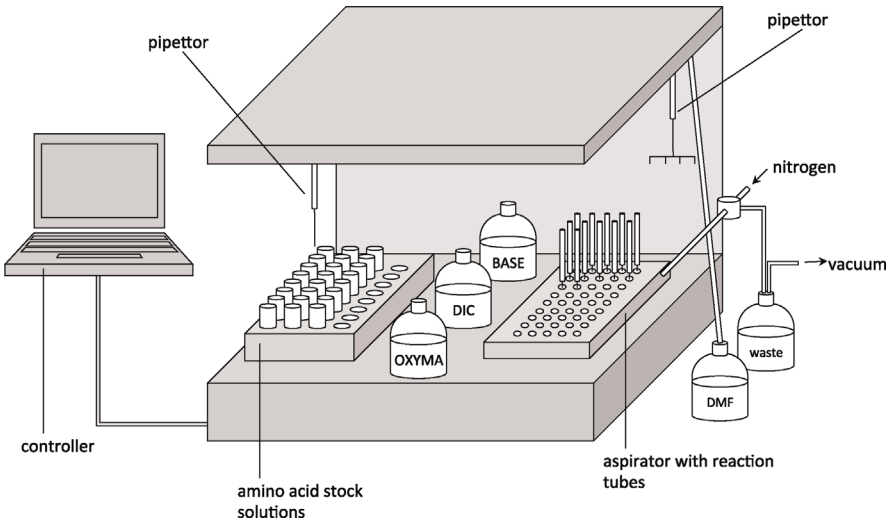


Figure 22.5 Peptide synthesizer for the parallel synthesis of multiple peptides.

prevent side reactions. Scavengers may be electron-rich aromatic compounds (thioanisole, cresol) and thiols (ethanedithiol) to remove cleavage products from the reaction mixture. As many thiols are strong-smelling and toxic, silanes like triisopropylsilane (TIS) are increasingly used for these purposes.

In many cases, it might be useful to have a fluorescent label on the peptide. These labels can be easily introduced when the peptide chain is still connected to the resin. By using protecting groups, which can be cleaved under different conditions, a specific side chain can be deblocked and used for labeling (e.g., a single lysine residue). N-terminal modifications may be introduced after cleavage of the last Fmoc group. The fluorescent dyes may be either added directly as active esters or activated *in situ*.

The consecutive development of solid phase peptide synthesis led in 1985 to multiple peptide synthesis, which means, the possibility to simultaneously synthesize multiple peptides with different lengths and sequences. The recurring cleavage, coupling, and wash steps allow an automation of the whole synthesis process (Figure 22.5).

For example, the synthesis may be conducted in polypropylene syringes with a frit bottom. Attached to a special fitting, reaction solutions may be aspirated by applying a vacuum after the reaction has taken place. During the reaction, drainage of the reaction solution is prevented by a slightly elevated nitrogen pressure. Addition of reagents and solution is realized by a computer-controlled pipettor arm. The desired peptide sequences and reaction conditions are collected in the controller, which calculates the amounts and concentrations of each compound. All compounds can then be filled in the designated container and the synthesizer conducts each step of the synthesis until the desired chain length is reached:

1. cleavage of the Fmoc group at the N-terminus of the peptide,
2. wash with DMF,
3. activation and coupling of the next amino acid,
4. wash with DMF,
5. second coupling of the amino acid (optional),
6. wash with DMF.

Cleavage of the peptides from the resin and the deprotection of all side chains is performed with TFA in the presence of scavenger compounds. For example, Boc- and *t*-butyl protecting groups are converted into *t*-butyl cations and *t*-butyl trifluoroacetate, respectively, by a S_N1 mechanism. Without addition of scavenger, these compounds could alkylate the 3′-position of tyrosine residues. Moreover, other amino acids like Trp and Met may be subject to alkylation. Another prominent issue during peptide synthesis is the oxidation of methionine. This becomes even worse when an automated peptide synthesizer is used, since aerial oxygen is sucked through the resin in each aspiration step. The fraction of oxidized methionine varies depending on the position of the residue in the sequence, but may be up to 50%. The closer to the C-terminus

Table 22.1 Common side products and side reactions during peptide synthesis and identification.

Side product/side reaction	Method of identification	Method of characterization
Scavenger (thioanisole, thiocresol)	HPLC	UV (photodiode array)
Salt	Amino acid analysis	NMR
Failure sequences	HPLC, CZE	MS [M–$M_{Amino\ acid}$]
Racemization and formation of diastereomers	HPLC, CZE	GC-amino acid analysis HPLC using chiral phase
Incomplete cleavage of side chain protecting groups, for example, Arg (Pmc)	HPLC	UV (photodiode array) MS: [M + 266]
Alkylation by *t*-butanol or incomplete cleavage of the *t*-butyl side chain protecting group	HPLC, MS	MS: [M + 57]
Oxidation of Met	HPLC	MS: [M + 16]
Deamidation from Gln, Asn	HPLC	MS: [M – 17]
Polymers	HPLC	MS: [2 M]
Addition of piperidine to Asp-Xaa sequences, often followed by elimination of water	HPLC	MS: [M + 67], [M – 18]
Sulfonation by Arg (Pmc) or Arg (Mtr)	HPLC, CZE	MS: [M + 81]

the methionine residue is located, the more often it is exposed to aerial oxygen and, thereby, the higher the fraction of oxidized methionine may be.

Many different problems may occur in the different steps of the peptide synthesis. This has to be taken into account when the synthesized peptide is analyzed (Table 22.1). Non-peptidic impurities may be caused by the remains of scavenger or adducts of scavenger and protecting groups. Peptidic contaminations are usually caused by false sequences, resulting from incomplete coupling reactions or deprotection of the N-terminus in one or more steps. Moreover, racemization during the coupling reaction may lead to diastereomers. Other impurities may be produced by modification of amino acid residues (alkylation), by incomplete deprotection of the side chains, or by oxidation of methionine side chains.

Usually, the analytics of the peptides (HPLC, UV, amino acid analysis, MS) are not limited by the amount of sample, as the yield of the peptide synthesis is on the milligram-to-gram scale.

22.2 Purity of Synthetic Peptides

The first task after cleavage of the peptide from the resin is directed towards the purity. For this purpose, reversed phase HPLC is applied as the standard method, usually using RP-18 as solid phase and acetonitrile/water/trifluoroacetic acid as mobile phase. In some cases it might be necessary to substitute trifluoroacetic acid with formic acid (e.g., to be compatible with subsequent ESI-MS analysis). The term "reversed phase" indicates that the solid phase is nonpolar, while the mobile phase is polar. To achieve narrow peaks and short retention times, a gradient elution is used. The composition of the mobile phase is continuously changed during the elution process. Most synthetic peptides may be analyzed using a gradient from 10% acetonitrile/90% water to 60% acetonitrile/40% water in 30 min. Trifluoroacetic acid is added to both solutions to fully protonate all peptides (0.08–0.11%, pH 2). It is used here because of its low absorbance at 214 nm. However, disadvantages of TFA are its high volatility and its toxicity. If the peptide is longer than 30 amino acids or very hydrophobic, 25% acetonitrile/75% water is recommended as the starting point of the elution gradient. *The maximum change of the eluent composition should not exceed 2% per minute.*

Chromatographic Separation Techniques, Chapter 10

If the synthesized peptides are strongly hydrophobic (e.g., transmembrane segments of proteins), analysis and separation will be difficult using the conditions described above. For such peptides, high amounts of formic acid in the eluent might be promising. A gradient ranging

from 60% formic acid/24% water/16% acetonitrile to 60% formic acid/4% water/36% acetonitrile is suggested in this case. Another option is to use columns with less hydrophobic material (e.g., *CN*, C_8, or C_4).

For the detection of peptides eluted from the column, UV-detectors with fixed wavelength or diode array detectors are commonly used. Eluents containing only water, acetonitrile, and trifluoroacetate allow detection at 214–220 nm, which is very close to the absorption maximum of the peptide bond. A photodiode array detector may additionally provide information about the identity of the products. (cf. Figure 22.3). When the eluents contain formic acid, the peptide products can only be detected at 280 nm. This wavelength corresponds to the absorption maximum of aromatic amino acids. Accordingly, peptides lacking any aromatic amino acid cannot be detected.

Another method which is gaining in importance for the analysis of synthesized peptides is capillary zone electrophoresis. This method is based on the separation of differentially charged peptides in an electric field. To optimize the separation of the synthesized peptides, the pH or ion strength of the buffer may be varied. As peptides often show a high pI value, the optimization at high pH values is often not successful. To overcome this problem, cation surfactants like hexadecyltrimethylammonium chloride may be added. The separation is now based on the differential distribution of the peptides between the aqueous phase and the ionic micellar pseudophase. Accordingly, this method is called "micellar electrokinetic capillary electrophoresis" and can be used to separate peptides that have a similar size, shape, charge, and hydrophobicity. As an example, the separation of four 17-mer peptides that differ only in one hydrophobic position is shown in Figure 22.6.

In some special cases, thin-layer chromatography (TLC) may be applied to analyze the purity of synthetic peptides. Whereas in most cases TLC is not suitable, some reactions can be monitored, for example, the modification of amino groups (amino groups may be detected using ninhydrin spray), formation of disulfide bonds (free –SH groups can be detected by Ellman's reagent), or phosphorylation of Ser, Thr, and Tyr.

If the amount of salt in the peptide is of importance, an amino acid analysis is performed. In such cases, the peptides are hydrolyzed again and the resulting amounts of each amino acid are compared to the theoretical amounts. Since Asn and Gln are converted into Asp and Glu during the hydrolysis process, the amounts of Asn and Asp (and Gln and Glu, respectively) may only be taken into account as the sum of both. A difference in the total amounts (resulting versus theoretical) representing 20–40% salt is not unusual, but is highly dependent on the amount of charged amino acids in the sequence. Moreover, trifluoroacetate salts show a higher salt content than hydrochlorides or acetates, as trifluoroacetate has a higher molecular weight. The substitution of trifluoroacetates by chlorides can be achieved by lyophilization from 0.1 N HCl solution. However, a disadvantage of the hydrochlorides is their hygroscopic character. When the peptides are intended to be used in animal studies, acetate salts are favorable. These can be obtained by several lyophilization steps from acetic acid. To verify that no trifluoroacetate is left, ^{19}F or ^{13}C NMR spectroscopy may be used, as the CF_3 group exhibits very characteristic signals.

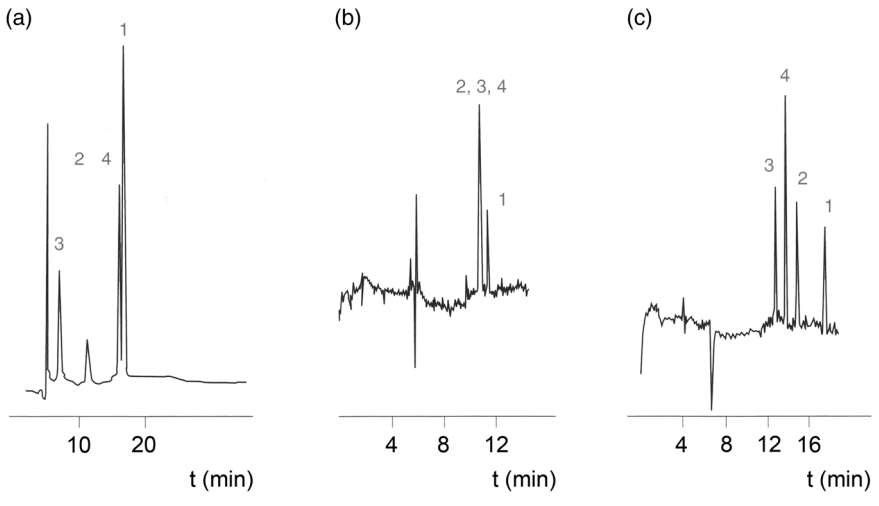


Figure 22.6 Separation of four different peptides by HPLC (a), capillary zone electrophoresis (b), and micellar electrokinetic capillary zone electrophoresis (c): YPSKLRHYIN**L**ITRQRY (1), YPSKLRHYIN**A**ITRQRY (2), YPSKLRHYIN**G**ITRQRY (3), and YPSKLRHYIN**I**ITRQRY (4).

Additionally, the amino acid analysis reveals information about the relative amount of amino acids, and which residues could have been coupled with low efficiency. Alanine is often used as reference amino acid because of its high abundance and because this residue is not modified either during synthesis or during hydrolysis and derivatization.

22.3 Characterization and Identity of Synthetic Peptides

When the chromatogram reveals several peaks the question arises as to which one corresponds to the desired peptide sequence and which side products can be found. When using a diode array detector and thereby recording a full UV spectrum of each peak, non-peptidic contaminants like scavengers or protecting groups can be identified directly by the absorption characteristics. The typically high absorption at 210–220 nm is missing for these compounds while aromatic absorption is over-pronounced. Side chain modifications of Tyr and Phe residues can also be identified based on the shift of the aromatic absorption peak. Additionally, protecting groups like Mtr (4-methoxy-2,3,6-trimethylbenzolsulfonyl, Figure 22.7b) that could not be cleaved can be detected in the UV spectrum. The N-terminal protecting group Fmoc can easily be identified by its typical maxima at 266, 289, and 300 nm. The molar extinction coefficients $\varepsilon_{266\ nm} = 17500$, $\varepsilon_{289\ nm} = 5800$, and $\varepsilon_{300\ nm} = 7800$ allow the quantification of the remaining Fmoc-groups on the peptide. The detection of two or more wavelengths also allows the identification of modified peptides with different absorption spectra. The desired product of an incompletely modified peptide can be distinguished from the unmodified one by the additional absorption of the modification (e.g., a Cy3 dye, absorbing at 552 nm, cf. Figure 22.8). This information can now be used to optimize the chromatographic conditions for the separation of the desired product by preparative HPLC.

Mass Spectrometry, Chapter 15

Today, the most important method for the characterization of synthetic peptides is mass spectrometry. Suitable spectrometers offer soft ionization techniques such as electrospray ionization or matrix-assisted laser desorption/ionization and are capable of detecting intact peptides. When using electrospray ionization, a series of charged $[M+nH]^{n+}$ ions is generated and subsequently detected in the mass spectrometer. Larger peptides are protonated at the N-terminus, on Lys and Arg residues, and therefore form multiple-charge ions. Figure 22.9a shows an ESI mass spectrum of a purified 16mer peptide amide. Besides the $[M+H]^+$ with $m/z =$ 1774.5, the twofold charged ion ($[M + 2H]^{2+}$, m/z 887.5) and the three-fold charged ion ($[M + 3H]^{3+}$, m/z 592.5) can be detected. The mass spectrum of the crude peptide is shown in Figure 22.9b. Several additional peaks can be detected, two of which represent common peptidic impurities. By calculating the mass difference between the $[M + H]^+$ signals or the $[M + 2H]^{2+}$ signals, mass differences of M − 156 amu and M + 266 amu can be obtained. The first represents a failed coupling of an Arg residue, the latter signal is caused by the complete peptide plus a Pmc protecting group still being attached to an Arg side chain (incomplete cleavage).

More difficult is the analysis of hydrophobic peptides using ESI-MS, as these peptides have only a few polar residues that can be protonated. Glu and Asp residues may lead to a negative net charge, which causes peptide fragmentation. As a consequence, M^{2+} and M^{3+} ions usually do not dominate the spectrum, which makes it difficult to evaluate the quality of the synthesized peptide. The highly fragmented spectra of hydrophobic peptides may be interpreted as follows:

1. search for M^{2+} and M^{3+},
2. search for failure sequences,
3. identify N- and C-terminal fragments,
4. identify failure sequences of N- and C-terminal fragments,
5. interpretation of the results.

Accordingly, the spectrum shown in Figure 22.10 corresponds to just one peptide, although the spectrum is quite complex and M^{2+} and M^{3+} are underrepresented. In addition to both ions, signals caused by fragmentation of the peptide are shown.

Another option to analyze hydrophobic or acidic peptides is to measure in negative mode. However, the detection of anions in ESI-MS is less sensitive and hard to optimize, especially for

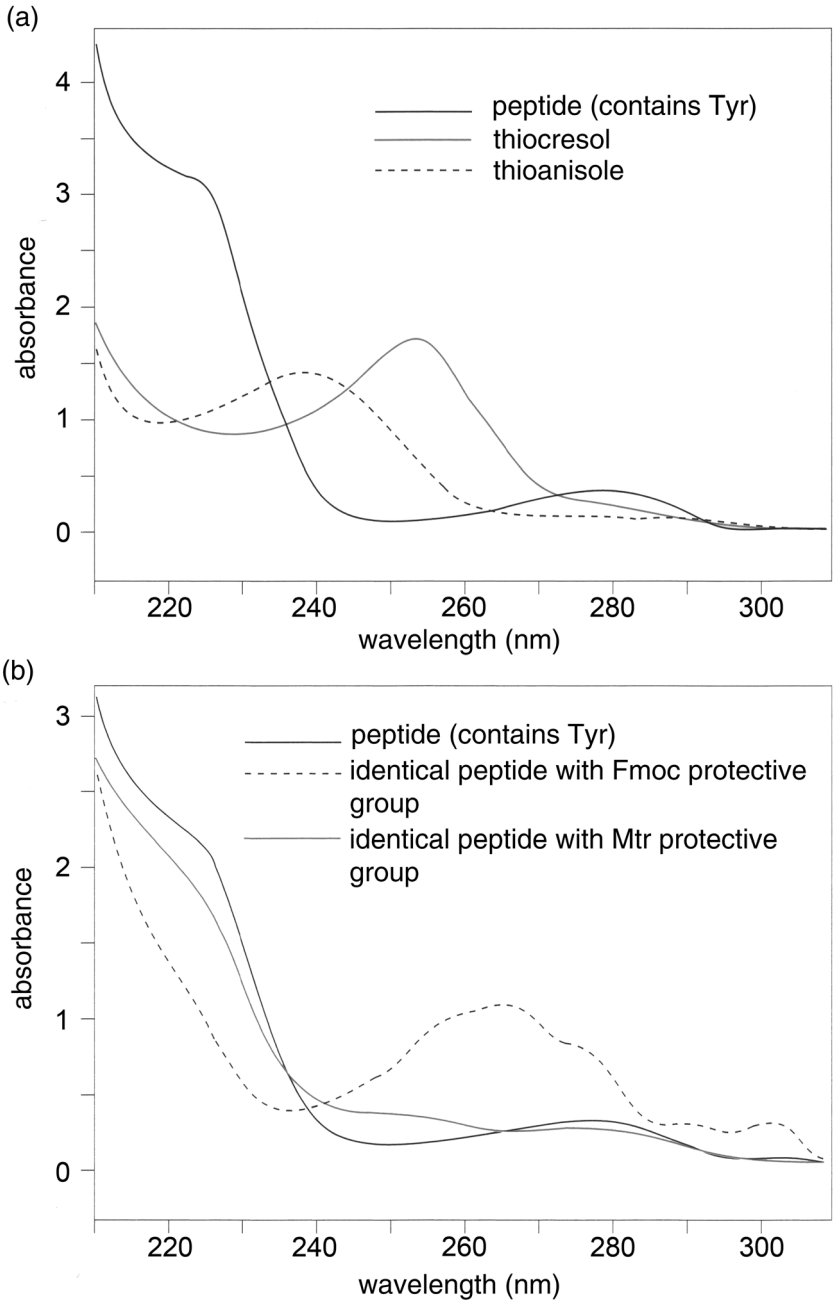


Figure 22.7 UV-Spectra obtained directly after HPLC by photodiode array detection. (a) Spectrum of a Tyr-containing peptide, thioanisole, and thiocresol. (b) Spectrum of an unprotected Tyr-containing peptide and the same peptide protected by a Fmoc or Mtr protecting group.

peptides. Taken together, mass spectrometry can reveal valuable information on the identity and purity of the synthesized peptide.

The synthesis of a peptide is successful when only one product with the correct mass is found in a chromatographically pure sample. If there are peaks in the mass spectrum that differ by the mass of one residue or one protecting group, failure sequences or incomplete cleavage will have to be considered.

Moreover, mass spectrometry can identify which segments of the peptide have been synthesized properly. Failure sequences may give hints as to which sequences may be hard to synthesize. Additionally, side products like peptides with oxidized methionine can be easily identified. Hence, HPLC and mass spectrometry is a powerful combination to evaluate the

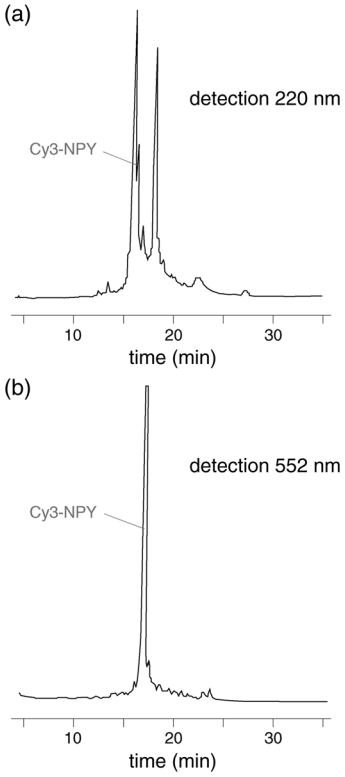


Figure 22.8 HPLC chromatogram of a peptide mixture. The neuropeptide hormone NPY (36mer) was labeled in solution with the fluorescent dye Cy3. (a) Detection at 220 nm shows all peptidic compounds in the sample. (b) Detection at 552 nm (corresponds with the absorption maximum of Cy3) leads to identification of the product peak at a retention time of 16.2 min.

quality of a synthesized peptide. Figure 22.11a displays the chromatogram of a crude synthetic 28-mer peptide, showing – in addition to the main product peak – a side product representing 18% of the total peptidic content of the sample. The corresponding mass spectrum reveals the correct mass of the main product and a second peak differing by M–199 amu from the main product, representing a missed coupling of the unnatural amino acid 12-aminododecanoic acid.

One side reaction that cannot be identified by mass spectrometry is the formation of diastereomers by racemization of single amino acids during the coupling reaction. Furthermore, Leu and Ile as well as Gln and Lys failure sequences may not be distinguished in the mass spectrum, as they lead to the same mass difference of −113 amu and −128 amu, respectively. Both problems may be solved by an amino acid analysis. However, amino acid analysis directly from the crude product may be challenging, and impurities <10% cannot be evaluated because of the error limit of the method. If the impurities are separated by HPLC before performing an amino acid analysis, Leu and Ile failure sequences may be identified.

Another option is the FT-ICR or Orbitrap mass spectrometers. Because of the high mass accuracy and the high resolution, Lys and Gln can be distinguished with these techniques. To investigate whether the product is enantiomerically clean, amino acid analysis in combination with gas phase chromatography or HPLC using a chiral solid phase may be applied. Using gas phase chromatography, the amino acids are modified after hydrolysis (e.g., as trifluoroacetyl amino acid *n*-propyl ester), separated by *Chirasil*-Val-capillaries and analyzed in a nitrogen detector. An internal D-amino acid standard allows the quantification and evaluation of racemization of single amino acids. This method is of high importance for the characterization of the enantiomeric purity of synthetic peptides and for the identification of peptidic natural products (e.g., antibiotics from fungi), which often contain D-amino acids.

To identify the position in the sequence of a modification, side product, or missed coupling, either N-terminal Edman sequencing or MS/MS-spectrometry can be performed. In the case of a failure sequence, the expected amino acid is missing and substituted by the next residue of the sequence. Modifications or side reactions will result in an empty cycle during Edman degradation, since the modified amino acid is not eluted within the normal retention times in the standard chromatogram. Peptide backbone modifications or β-amino acids lead to abortion of the Edman degradation at this position.

MS/MS is the induced fragmentation of peptides in the mass spectrometer, usually caused by collision with inert gas molecules. As described above, the fragmentation products can yield information about the location of a modification.

22.4 Characterization of the Structure of Synthetic Peptides

One of the most rapid methods to evaluate the conformation of a synthetic peptide is circular dichroism (CD) spectroscopy. Right-handed α-helical peptides reveal CD spectra with a negative Cotton effect (CD bands are denominated as Cotton effect) at $\lambda = 222$ nm ($n \rightarrow \pi^*$-transition), a positive CD band at 192 nm, and a negative band at 207 nm. Both short-wavelength bands correspond to the carbonyl $\pi \rightarrow \pi^*$ transition, which is split into two components. Peptides showing a minimum between 215 and 220 nm ($n \rightarrow \pi^*$-transition) and a maximum at 195 nm ($n \rightarrow \pi^*$-transition) are folded in an antiparallel β-sheet conformation. Beyond this, unstructured peptides also show characteristic CD spectra (Figure 22.12). CD spectra are interpreted by comparison with standard peptides with known conformation. In the case of cyclic peptides, CD spectroscopy can reveal which cyclization position is better suited to stabilizing a certain secondary structure. Figure 22.13 shows the CD spectra of two cyclic dodecamer peptides, which differ solely in the configuration of one amino acid that is part of the bridge. The L-configured bridge stabilizes the desired α-helical structure much better than the D-configured bridge.

Circular Dichroism, Section 7.7.2

CD spectroscopy is predominantly suited for the comparison of homologous peptides, for the optimization of NMR conditions (e.g., the identification of the best solvents), for evaluation of the stability of conformations (pH titration, different temperatures, different solvents), and for the characterization of rigid structures.

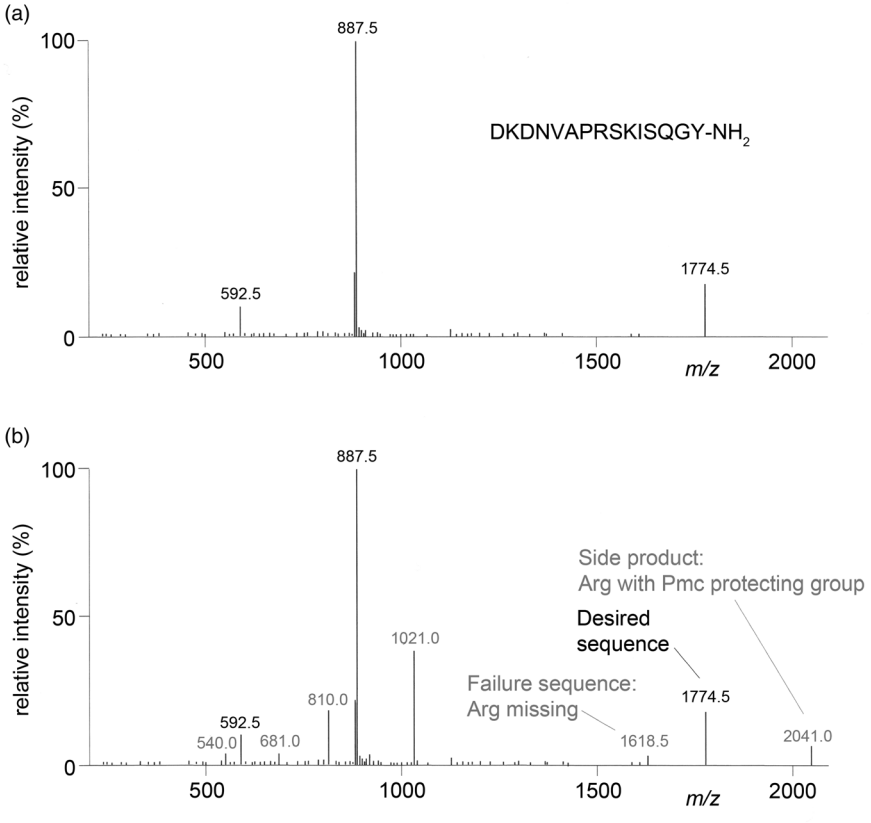


Figure 22.9 Electrospray mass spectrometry of a purified 16mer peptide amide (M = 1773.5 amu) (a) and of the crude product (b), which still contains contaminants with a mass of M = 1617.5 amu (M–156 amu) and 2040.5 amu (M + 266 amu), respectively. The mass shifts indicate incomplete coupling of Arg and an incomplete cleavage of the arginine side chain protecting group Pmc.

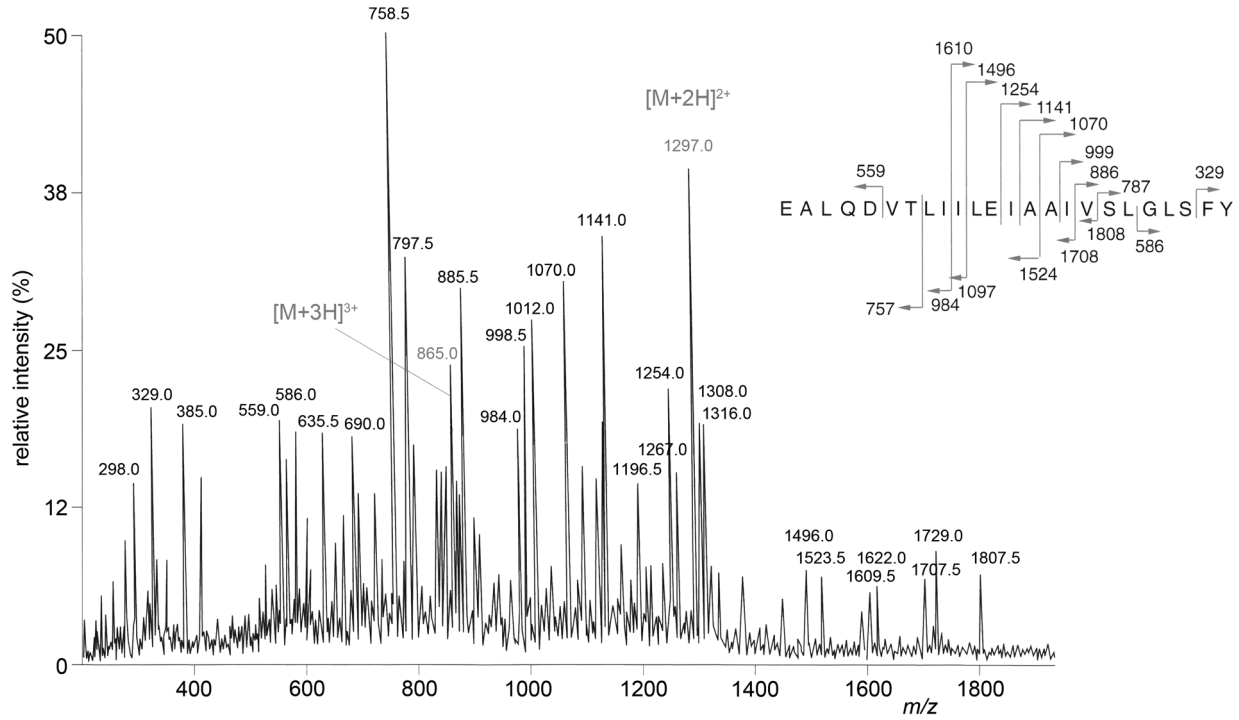


Figure 22.10 Electrospray mass spectrum of a hydrophobic transmembrane peptide fragment (24mer) lacking any basic amino acid. The spectrum displays an $[M + 2H]^{2+}$ peak at 1297 amu and an $[M + 3H]^{3+}$ peak at 865 amu. The fragmentation pattern indicates that the peptide was synthesized with the desired sequence.

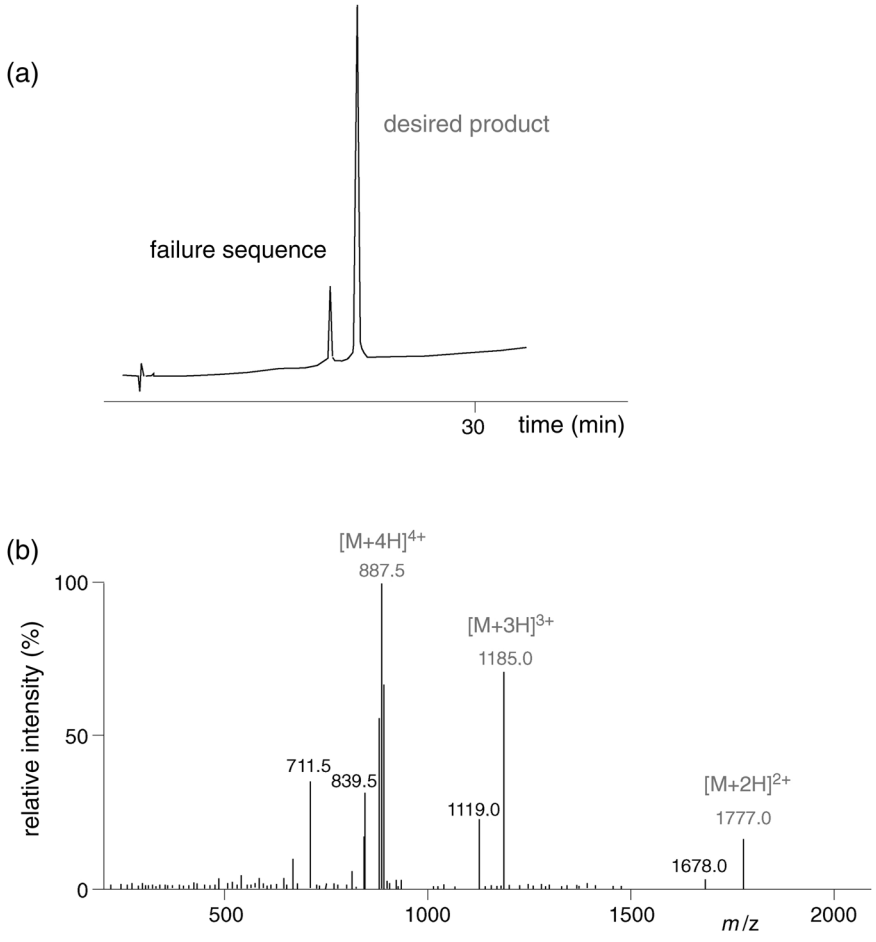


Figure 22.11 The combination of HPLC (a) and electrospray mass spectrometry (b) allows the identification and characterization of the desired product (3552 amu, 82%) and a failure sequence (3354 amu, 18%).

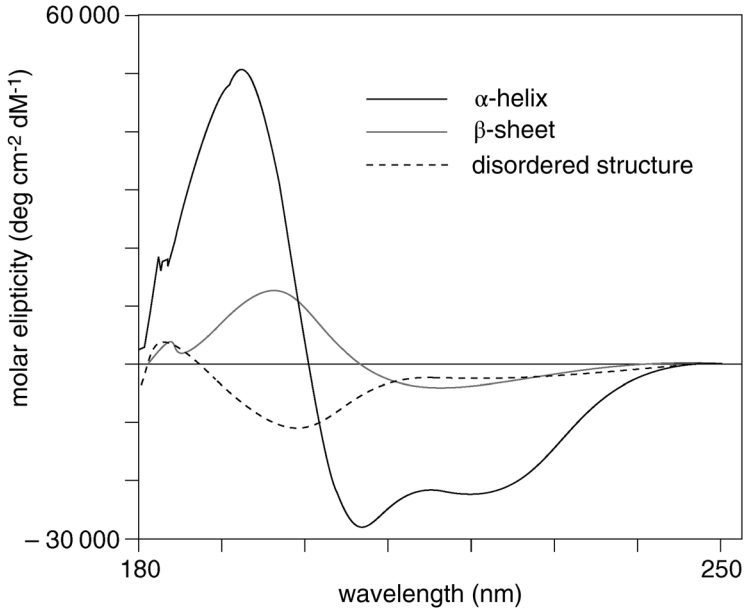


Figure 22.12 Circular dichroism spectra of 15mer peptides with typical α-helical conformation, β-sheet structure, and no defined structure, obtained in 10 mM phosphate buffer (pH 7) and trifluoroethanol (2 : 1); dM = decimolar.

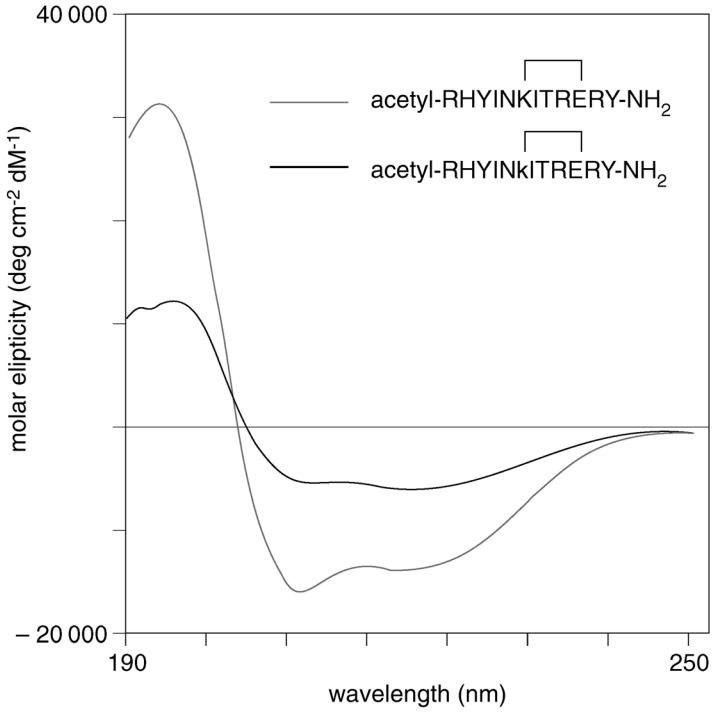


Figure 22.13 Circular dichroism spectra of two cyclic peptides (dM: decimolar). The peptide containing an L-amino acid (K) shows a higher content of α-helical structure than the one containing the corresponding D-amino acid (k).

More accurate, but also much more complex, methods for the characterization of the structure of synthetic peptides are 2D and 3D NMR spectroscopy or X-ray crystallography. At best, both methods can reveal the exact spatial structure or at least distinguish rigid from flexible structures.

NMR, Chapter 18
X-Ray Structure Analysis, Chapter 21

22.5 Analytics of Peptide Libraries

Mixtures of synthetic peptides, so-called peptide libraries, are used in the development of new compounds for plant protection or therapeutic purposes. Peptide libraries are mixtures of n^m peptides, where n denotes the number of varied amino acids at one single position and m represents the number of varied positions (in the sequence denominated as "X"). A hexapeptide library that contains each of the 20 naturally occurring amino acids on each position consists therefore of $20^6 = 64\,000\,000$ individual peptides. A hexapeptide library with only three variable positions and only ten different amino acids per variable position contains $10^3 = 1000$ individual peptides. These libraries are generated by coupling of peptide mixtures or by the so-called *divide–couple–combine* method. In the latter case, the polymer is distributed in m different vessels, in which the coupling takes place. After coupling, the polymer from all vessels is mixed and distributed again in m vessels (Figure 22.14). Assuming a 100% coupling efficiency, each polymer bead carries only one individual peptide sequence. The active peptides are then evaluated in bioassays. When libraries contain free peptides or have been generated by coupling of peptide mixtures, identification of the active sequence is performed by synthetic partial libraries, which contain fewer variable positions (Figure 22.14). If polymer-bound libraries are used in the bioassay, the active sequence may be identified directly from the resin. However, the characterization remains complex.

Edman degradation is well-suited for the analysis of a peptide library because of its high sensitivity. This allows sequencing of a peptide directly from one bead of the resin and therefore leads to the active substance, if the bioassay was performed on-bead. In addition, the so-called pool sequencing has been established. The complete mixture is subjected to Edman degradation. A defined position can be distinguished from variable positions by the eluted amino acid in the chromatogram: 50 pmol of a tetrapeptide library with the sequence LXTX (X = A,G,L,I) consists of 16 peptides. In the first step of the Edman degradation, only L is seen in the chromatogram, in the second step A, G, L, and I (1 : 1 : 1 : 1) are detected, in the third step solely T, and in the fourth step again A, G, I, and L (1 : 1 : 1 : 1) are detected. Thereby, at least the

Amino Acid Sequence Analysis, Chapter 14

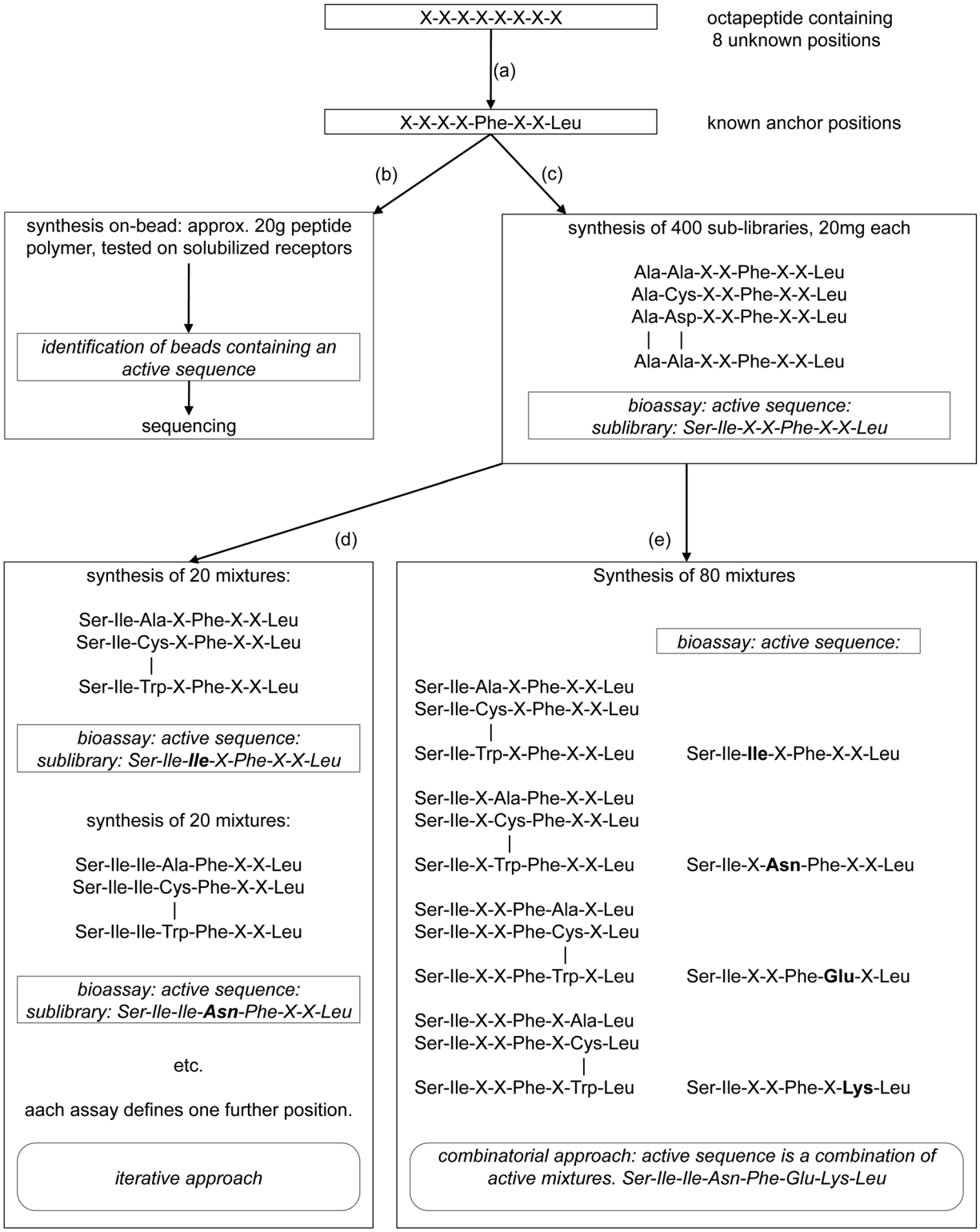
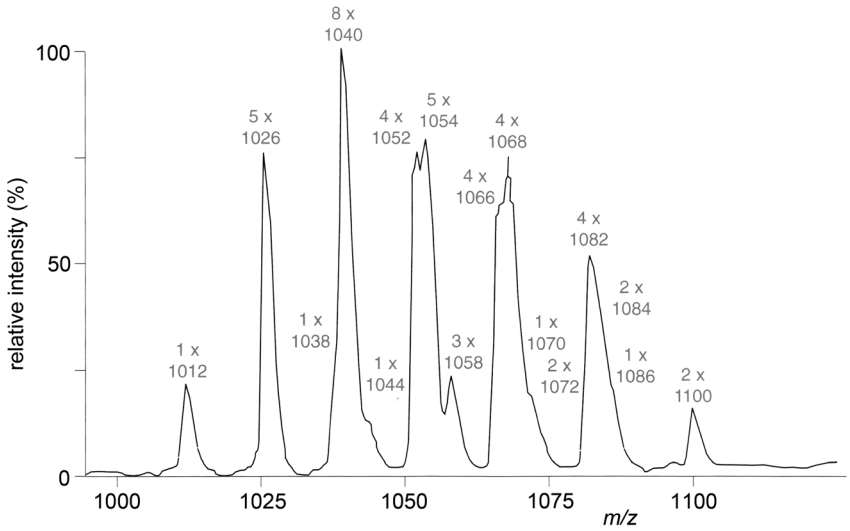


Figure 22.14 An approach to identifying a biologically active substance from a mixture using the example of the CTL-epitope of ovalbumin 258–276 (CTL, cytotoxic T-cell lymphocytes). Two anchor positions are known (a). An 8-mer peptide library (X = any of the 20 naturally occurring amino acids) may be synthesized according to the divide–couple–combine method (b) or by the synthesis of sub-libraries (c). For the latter, 400 sub-libraries containing two fixed and four variable positions must be synthesized. The variable positions may be identified either sequentially or by combination (e). In both cases, another 80 peptide mixtures are needed. Source: Jung, G. and Beck-Sickinger, A.G. (1992) *Angew. Chem.*, **104**, 375–391.



Pos.	A Ala	R Arg	N Asn	D Asp	E Glu	Q Gln	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	Y Tyr	V Val
1	0.3	42950.0	42861.0	42790.0	42738.0	0.5	0.0	42981.0	20.0	**2860.2**	42859.0	42796.0	42953.0	42948.0	42835.0	280.5	42951.0	42830.0
2	0.0	0.0	**4007.9**	411.1	42798.0	42892.0	0.0	0.1	0.0	231.3	0.0	42827.0	42797.0	0.7	42769.0	210.1	42936.0	42859.0
3	0.1	42797.0	382.5	47.3	42797.0	42831.0	0.0	32.5	42951.0	42900.0	42952.0	42858.0	42744.0	42738.0	42887.0	120.0	**6709.4**	42771.0
4	0.3	**1327.0**	141.7	42844.0	7.0	42894.0	0.0	5.0	42830.0	0.1	42802.0	42772.0	47.1	0.0	42768.0	117.6	308.7	42921.0
5	0.0	564.5	63.6	42954.0	42995.0	42986.0	0.0	42830.0	42807.0	0.4	0.0	42829.0	**6749.1**	0.1	42860.0	81.0	43008.0	42772.0
6	0.7	152.9	37.0	42873.0	**1474.5**	42809.0	42740.0	42863.0	**1019.1**	42888.0	42959.0	42982.0	132.3	0.0	**532.3**	**1074.4**	42748.0	42923.0
7	0.0	48.7	**1571.3**	151.5	463.6	**1695.5**	2.0	42738.0	42877.0	42740.0	**1546.5**	42786.0	42967.0	0.3	42960.0	42745.0	42953.0	42928.0
8	0.0	47.1	92.5	42777.0	42978.0	64.1	0.6	42830.0	**456.2**	**416.8**	45.2	**695.3**	42744.0	42917.0	42918.0	55.6	42739.0	**973.0**
9	0.0	42748.0	42806.0	42949.0	42800.0	42892.0	0.0	0.8	0.0	0.2	0.0	42964.0	36.2	0.7	0.9	46.4	42796.0	42905.0
10	0.1	43000.0	42745.0	0.0	42860.0	6.0	0.0	0.8	42828.0	42769.0	42856.0	0.0	0.0	0.1	1.0	39.2	0.2	42982.0

Figure 22.15 Characterization of a peptide library by ESI mass spectrometry and pool sequencing. The synthetically made octapeptide library LNYRFX$_1$X$_2$X$_3$ (X$_1$ = T, I, E, S; X$_2$ = N, Q, K; X$_3$ = L, M, I, V) contains 48 peptides in total. The mass distribution pattern can be evaluated by ESI-MS. Source: Stevanovic, S. *et al.* (1993) *Bioorg. Med. Chem. Lett.*, **3**, 431–436.

quality of a library can be estimated. Another possible way to analyze peptide libraries is by mass spectrometry. If the composition of the library is known, the theoretical masses can be calculated (Figure 22.15). Comparison between the theoretical mass distribution and experimentally determined mass distribution gives information about the quality of the library. Partial modifications, which appear on all peptides of the library, result in a second identical but shifted mass distribution pattern. Similarly, side reactions from the synthesis can also be identified and characterized. However, to date there is no way to prove that all theoretically possible sequences are actually included in the library, that all peptides are equally distributed. Hence, the identification of missing sequences is also not yet possible.

Mass Spectrometry, Chapter 15

Further Reading

Atherton, E. and Sheppard, R.C. (1989) *Solid-Phase Synthesis – A Practical Approach*, Oxford University Press, Oxford.

Bettio, A. and Beck-Sickinger, A.G. (2002) Biophysical methods to study peptide protein interaction. *Biopolymers*, **60**, 420–437.

Chan, W.C. and White, P.D. (2000) *Fmoc Solid Phase Peptide Synthesis – A Practical Approach*, Oxford University Press, Oxford.

Jones, C., Mulloy, B., and Thomas, A.H. (eds) (1994) *Microscopy, Optical Spectroscopy and Macroscopic Techniques*, Methods in Molecular Biology, vol. 22, Humana Press, Totowa, NJ.

Koglin, N., Lang, M., Rennert, R., and Beck-Sickinger, A.G. (2003) Facile and selective nanoscale labeling of peptides in solution using photolabile protecting groups. *J. Med. Chem.*, **46**, 4369–4372.

Ahrens, V., Bellmann-Sickert, K., and Beck-Sickinger, A.G. Peptides and peptide conjugates: therapeutics on the upward path. *Future Med. Chem.*, **4**, 1567–1586.

Bellmann-Sickert, K. and Beck-Sickinger, A.G. (2010) Peptide drugs to target G-protein coupled receptors. *Trends Pharm. Sci.*, **31**, 434–441.

Carbohydrate Analysis

23

Peter Hermentin

Feuerdornweg. 8, 35041 Marburg, Germany

The development of recombinant glycoproteins has in recent years led to a surge of scientific, pharmaceutical, and regulatory interest in the structure and function of the carbohydrates expressed on these glycoproteins.

Glycoproteins are biopolymers that consist of a polypeptide backbone with carbohydrate side chains that are glycosidically linked to the amino acid residues of asparagine, serine, or threonine of the peptide backbone. The glycoproteins comprise most plasma proteins and numerous cytokines (hormonal messengers) as well as cell surface proteins. Nowadays, some glycoproteins are gained via genetic engineering of mammalian cells and are used for the treatment of various diseases (Table 23.1).

Most recombinant glycoproteins were found to exhibit *in vivo* efficacy only when they were glycosylated and their glycosylation (the type and nature of their carbohydrate side chains) resembled those of the natural human glycoprotein counterpart. Glycosylation analysis of therapeutic glycoproteins is therefore indispensable for the assessment of their biological activity, *in vivo* half-life, pharmacological safety, and batch consistency.

Table 23.1 Examples of therapeutic glycoproteins.

	Function	Indication
Erythropoietin (EPO)	Formation of red blood cells	Anemias
Granulocyte/macrophage colony stimulating factor (GM-CSF)	Formation of white blood cells	Leukemias
Cetuximab (Erbitux®)	A monoclonal antibody against EGF-R (human epidermal growth factor receptor)	Colorectal carcinoma; squamous cell carcinoma of head and neck
Trastuzumab (Herceptin®)	A monoclonal antibody against HER2 (human epidermal growth factor receptor 2)	Breast cancer
Bevacizumab (Avastin®)	A monoclonal antibody against VEGF (vascular endothelial growth factor)	Various cancers
Tissue plasminogen activator (t-PA)	Regulation of fibrinolysis	Heart attack, stroke
Antithrombin III (ATIII)	Anticoagulant	ATIII deficiency
Factor VIII	Blood coagulation	Hemophilia A
Factor XIII	Blood coagulation	Wound healing
Follicle-stimulating hormone (FSH)	Oocyte maturation	Infertility
Interferon-β (IFN-β)	Inhibition of proliferation	Multiple sclerosis

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

Today, one is committed to replace the very demanding methods of carbohydrate analysis (which require a high level of expertise in HPLC, electrophoresis, mass spectrometry, and high-resolution ^{1}H NMR spectroscopy and the corresponding instrumental infrastructure) with routine, expeditious, simple, reliable, cost-effective, and efficient standardized chromatographic methods.

In terms of protein glycosylation, we are in this chapter only interested in the carbohydrate-specific modifications of a glycoprotein. We tacitly assume that the glycoprotein to be analyzed is available in pure form and therefore will not deal with the various aspects of protein and glycoprotein purification.

Furthermore, we shall in the following focus on the glycosylation of mammalian, respectively, human glycoproteins, and will only rarely use examples from other species. This seems permissible, as the methods of carbohydrate analysis described below of (perhaps recombinantly produced) human glycoproteins fully apply to glycoproteins from other sources (e.g., from plant cells, yeast, bacteria, or viruses).

As the analysis of oligosaccharide side chains of glycoproteins is a rather complex and stereochemically demanding task, we shall first explain some of the fundamental concepts of the stereochemistry of carbohydrates. The focus is on, on the one hand, the variety of possible stereoisomers and, on the other hand, on the various aspects of the glycosidic bond.

23.1 General Stereochemical Basics

Carbohydrates are polyhydroxy aldehydes or polyhydroxy ketones. They have multiple reactive groups and can be combined (each with the elimination of a water molecule) in quite manifold manners. Whereas two molecules of the same amino acid can combine to give a single dipeptide, the combination of two glucose molecules may result in 11 different glucose disaccharides (Table 23.2). Two different hexoses (e.g., D-glucose and D-galactose) may be combined to a total of $2^4 = 16$ different disaccharides (disregarding the possible variation of the anomeric linkage, α-α, α-β, β-α and β-β), and three different hexoses may already result in $3^4 = 81$ different trisaccharides.

The analysis of oligosaccharides thus not only requires the determination of the *monosaccharide sequence* but also for each of the monosaccharide units the assignment of its anomeric configuration and its respective linkage direction (Section 23.1.5). The relationship of the monosaccharides with each other is best derived from their synthesis scheme (Figure 23.1).

23.1.1 The Series of D-Sugars

Sugars exhibiting at their asymmetric carbon atom that is farthest from the carbonyl group the same configuration as D-glyceraldehyde (Figure 23.1) are referred to as D-sugars (for historical reasons). In the Fischer projection formula of D-glyceraldehyde, the OH group at the asymmetric

Table 23.2 Glucose disaccharides.

Abbreviation	Name	Abbreviation	Name	Abbreviation	Name
Glc1α-1αGlc	α,α-Trehalose	Glc1β-1αGlc	β,α-Trehalose (Neo-trehalose)	Glc1β-1βGlc	β,β-Trehalose (isotrehalose)
Glc1α-2Glc	Kojibiose	Glc1β-2Glc	Sophorose		
Glc1α-3Glc	Nigerose	Glc1β-3Glc	Laminaribiose		
Glc1α-4Glc	Maltose	Glc1β-4Glc	Cellobiose		
Glc1α-6Glc	Isomaltose	Glc1β-6Glc	Gentiobiose		

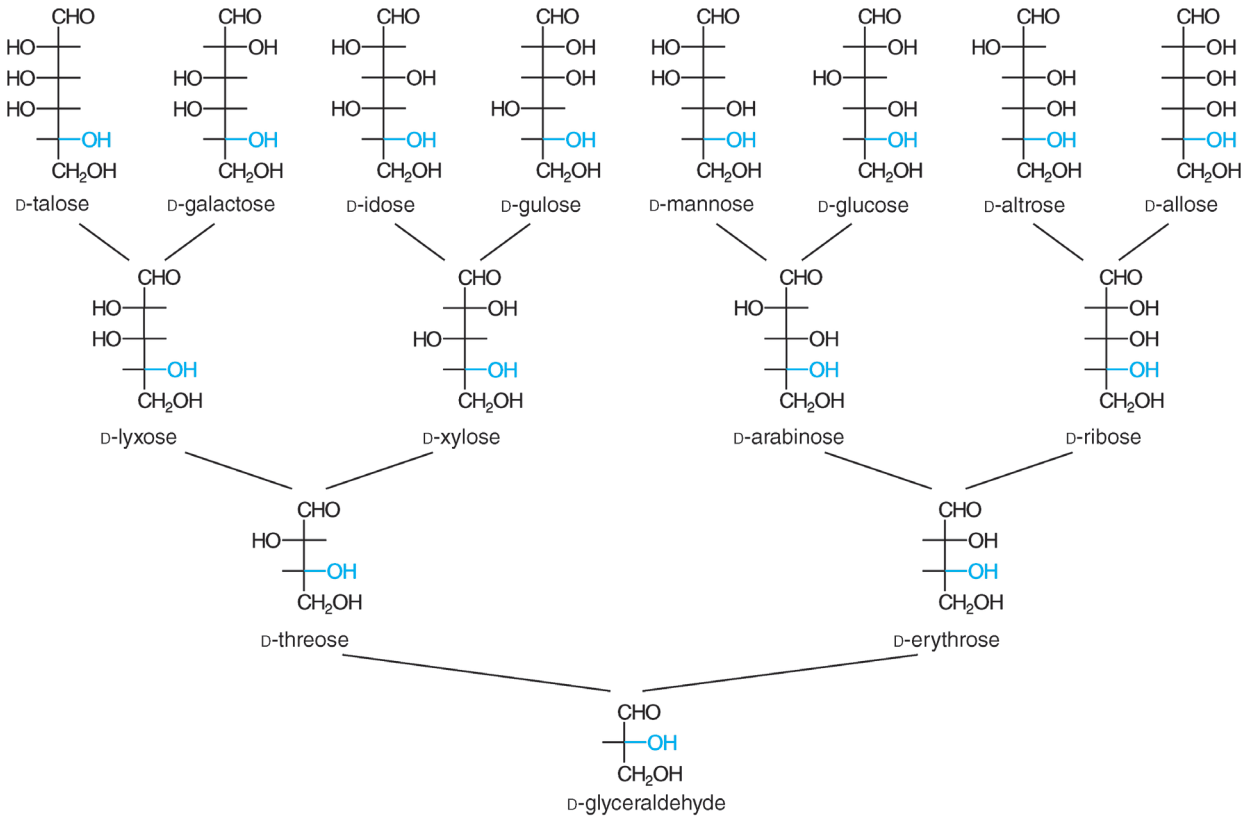


Figure 23.1 Composition of the series of D-sugars starting from D-glyceraldehyde.

carbon atom (C2) is by definition located on the right-hand side. In his basic stereochemical work, Emil Fischer had made this simple suggestion and (based on this assumption) then drawn the stereochemical system of D-sugars by synthesis. In his pioneering studies he succeeded in 1890 in the assignment of the relative configuration at all four chiral centers of the open-chain glucose. For this he was awarded in 1902 the Nobel Prize in Chemistry.

The initial situation, in which Emil Fischer had the OH group at the asymmetric carbon atom of D-glyceraldehyde arbitrarily drawn on the right side, was fortunately proved to be correct in later (X-ray crystallographic) studies.

23.1.2 Stereochemistry of D-Glucose

The D-hexoses are shown in more detail below because of their enormous biochemical importance as building blocks, using D-glucose as an example.

In Fischer's projection formula, the OH groups on the asymmetric carbon atoms (C2 to C5) of D-glucose (Figure 23.2a) are arranged in the order of "ta tii ta ta" (memory aid). However, hexoses are by their nature not open-chain structures but form six-membered ring structures, the pyranoses. Thereby, the OH group at C5, which is favored for steric reasons, reacts with the C1-aldehyde group to give one of the two possible cyclic hemiacetals, α-D-glucopyranose or β-D-glucopyranose, which are interconvertible via an equilibrium reaction. Figure 23.2b shows these reactions in Haworth's projection formula.

However, Haworth's projection formula of D-glucopyranose also represents the stereochemistry of this sugar block too simplistically and is therefore misleading. Due to the tetrahedral arrangement of the ligands at the various asymmetric carbon atoms, D-glucopyranose is actually angled and on average in the lowest energy chair conformation (Figure 23.2c).

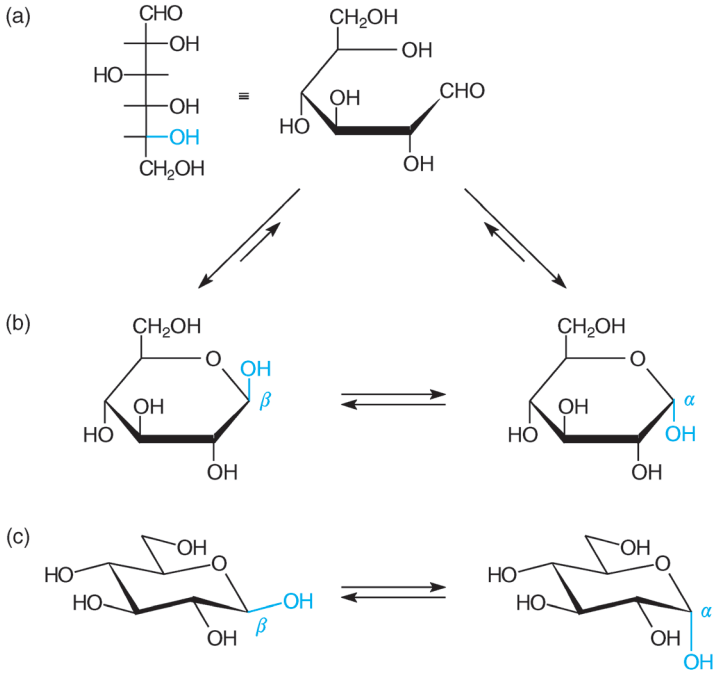


Figure 23.2 Projection formulas for D-glucose. (a) Fischer projection formula (ring-opened aldehyde structure), hint (as memory aid): ta tii ta ta. (b) Haworth projection formula, hint: starting from C5, large substituents are alternating up down up down. The OH-group at C1 (the anomeric center) is either up (β-configuration) or down (α-configuration). (c) Chair form. The energetically most favorable chair conformation of D-glucopyranose is obtained from the Haworth projection by flipping C4 up and C1 down. This ^4C conformation is, with respect to all other possible conformations of D-glucopyranose, energetically favored because the major ligands (namely, CH_2OH at C5 and the OH-groups at C4, C3, and C2) are all equatorial and therefore cause the least steric interaction.

23.1.3 Important Monosaccharide Building Blocks

Figure 23.3 summarizes (for our analytical consideration) the most important representatives of D- and L-hexoses. These are part of the *N*-glycosidic sugar chains (N-glycans) of mammalian cell glycoproteins, which play an important role in cell recognition and cell differentiation. *N*-Acetylneuraminic acid (sialic acid), a nona-sugar of the D-series, has an important signaling function as a terminal building block of *N*-glycans. Pentoses and hexofuranoses are, with respect to the analysis of mammalian cell glycoproteins, without importance, since they do not occur on them.

23.1.4 The Series of L-Sugars

The series of L-sugars is built up in analogy to the series of D-sugars, starting with L-glyceraldehyde. The glyceraldehyde-derived OH group in Fischer's projection formula for the L-sugars is on the left. The projective position of the OH group of an L-sugar (e.g., L-glucose) behaves, compared to the corresponding D-sugar (in our case D-glucose), as image and mirror image (Figure 23.4).

However, most naturally occurring sugars belong to the D-series (except for L-fucose). The examples described in the following sections are limited to the sugars of the D-series.

23.1.5 The Glycosidic Bond

The glycosidic bond of a sugar refers to the anomeric linkage to a sugar or alcohol residue. It is either α- or β-glycosidic in nature and one may (if the sugar is in solution) expect free rotation around the C—O—C single bond. In reality, however, the glycosidic bond is conformationally quite stable; free rotation around the single bond is in fact blocked owing to both steric (steric demand of neighboring sugar residues) and electronic factors (exo-anomeric effect, see below).

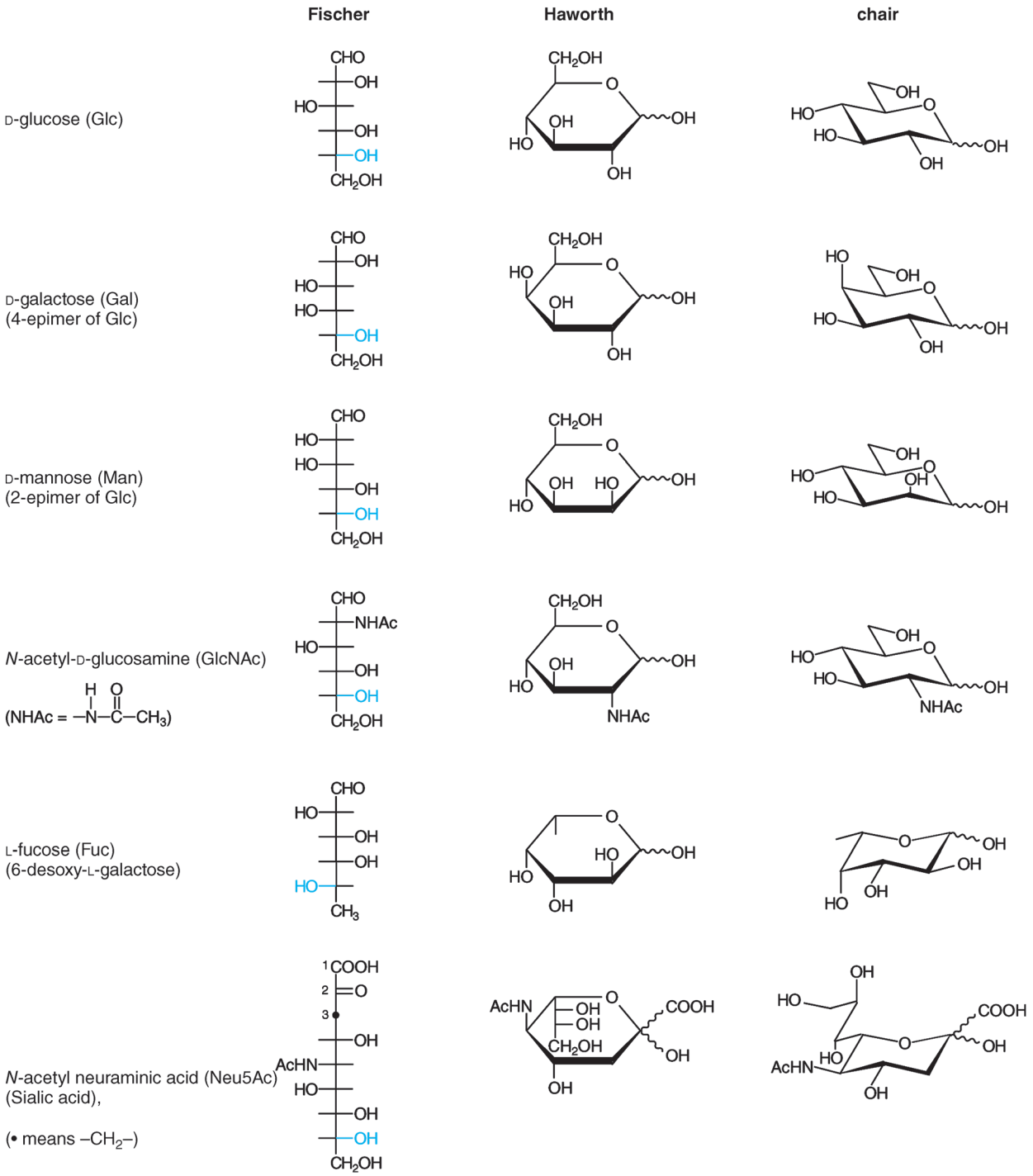


Figure 23.3 Important monosaccharide units of mammalian cell glycoprotein.

The Anomeric Configuration A change in the anomeric configuration (e.g., from α to β) results in elemental stereochemical changes of the respective di-, tri-, or oligosaccharide structure, as exemplified with starch and cellulose (Figure 23.5). These two glucose polymers have, as a result of α1,4- or β1,4-linkage of glucose units, a completely different spatial arrangement. This fact, proven by means of ^{1}H NMR spectroscopy (Section 23.3.5), can be explained by the so-called exo-anomeric effect, which is described below.

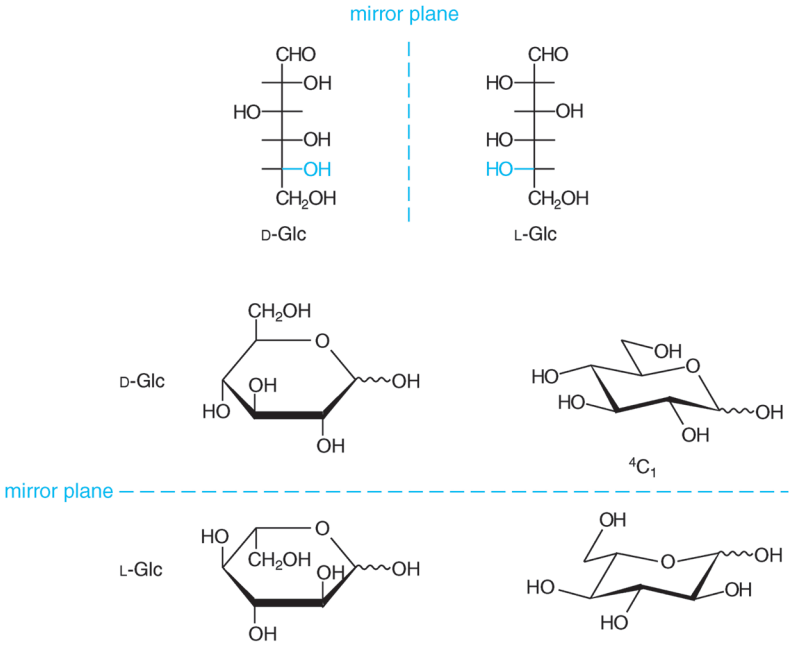


Figure 23.4 Mirror image isomerism of D-glucose and L-glucose.

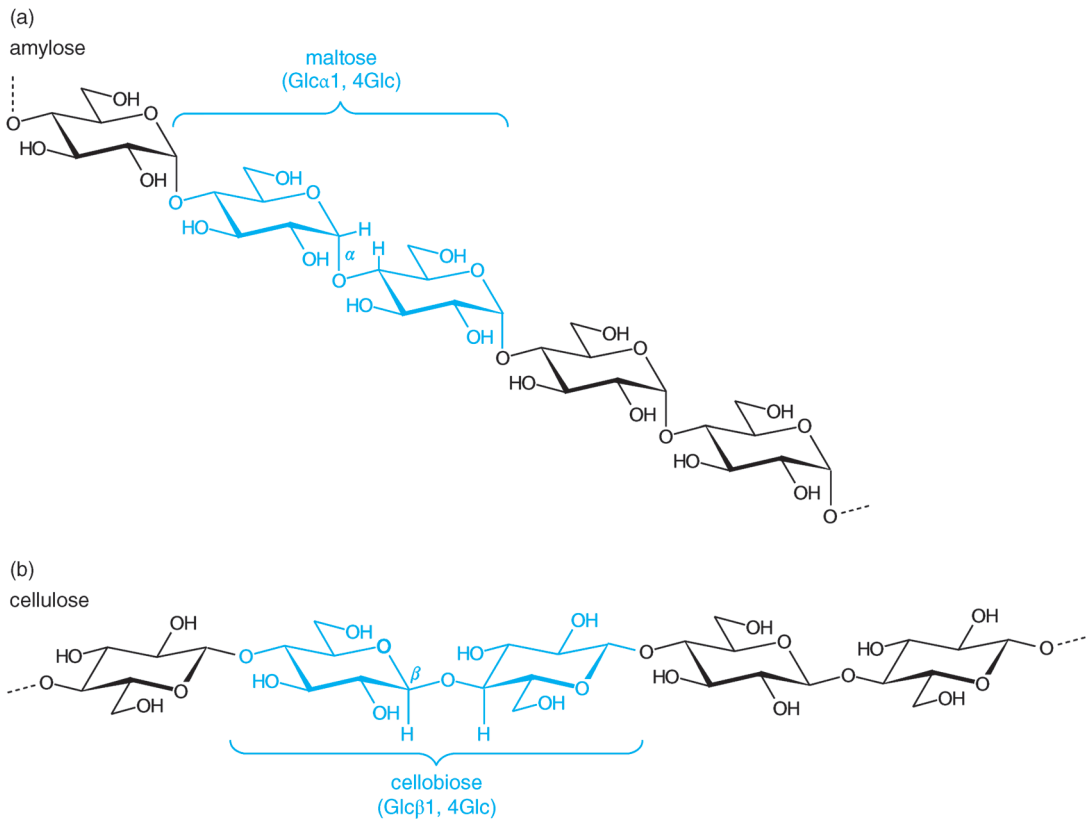


Figure 23.5 The anomeric configuration: (a) Amylose with maltose (Glcα1,4Glc) as repetitive basic building block. (b) Cellulose with cellobiose (Glcβ1,4Glc) as repetitive basic building block. As may be seen, the α-glycosidically linked D-glucopyranose (a) may be pushed along the amylose chain and brought into projective cover with all successive glucose units. (In reality, amylose is not elongated because of the α-glycosidic bond, but helically wound and therefore allows the formation of inclusion compounds; for example, the inclusion of iodine yields an intense blue staining.) If, however, the β-glycosidically linked D-glucopyranose is analogously moved along the cellulose chain (b), this can be brought into projective cover only with every second glucose unit.

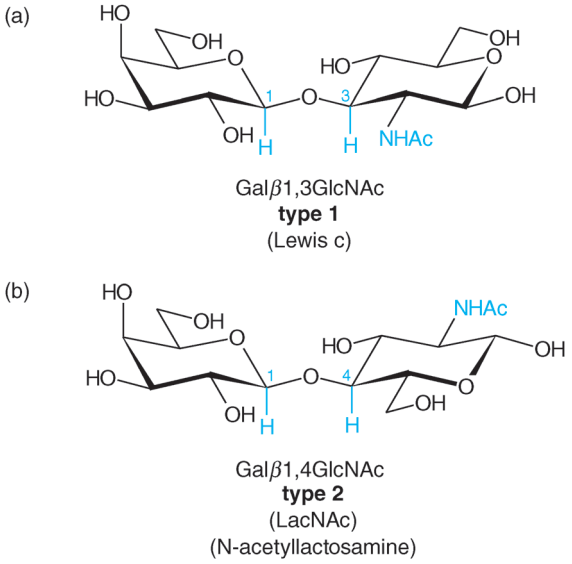


Figure 23.6 Gal-GlcNAc binding of type 1 (a) and type 2 (b).

The Linkage Direction A change in linkage direction leads to a disaccharide molecule with suddenly changed properties. For example, the blood group determinants of the ABH(0) system of type 1 exhibit a fundamental Galβ1,3GlcNAc building block, while the corresponding determinants of type 2 carry Galβ1,4GlcNAc as the basic structure (Figure 23.6). This stereochemical change is particularly visible from the position of the *N*-acetyl group, which in the Galβ1,3GlcNAc structure faces the viewer, so to speak, while in the corresponding Galβ1,4GlcNAc isomer it looks away from the viewer, that is, it is located behind the plane of the paper. Enzymes, antibodies, and lectins clearly can differentiate between such structural changes.

The Exo-anomeric Effect As can be deduced from stereochemical considerations, the rotation about the glycosidic bond is not free, due to the spatial claim of neighboring groups, but hampered significantly.

An oligosaccharide determinant cannot move freely, but takes, due to steric and electronic interaction, an energetically preferred conformation (exo-anomeric effect), which can be measured by ^{1}H NMR spectroscopy (cf. Section 23.3.5).

Electronic considerations likewise result in a spatial preference of the glycosidic bond: The four ligands of the ring oxygen (C1 and C5 as well as the two free pairs of electrons) are arranged tetrahedrally. While the O—C1 and O—C5 bonds are integrated into the chair conformation of the pyranose ring, the two lone pairs of electrons of the ring oxygen atom extend away from the pyranose ring in the axial and equatorial direction (Figure 23.7).

In the presence of a β-glycosidic linkage, the glycosidic O-atom points away from C1 of the pyranose ring in the equatorial direction towards the neighboring group. Thereby, the two free-electron pairs of the glycosidic O-atom enter a dipole–dipole interaction with the two lone electron pairs of the ring O-atom – in such a way that the dipole–dipole interaction energy is minimized. Consequently, the glycosidic residue R acquires a unique preference direction – this property is called the exo-anomeric effect. In this energetically favored conformation of the glycosidic bond, the anomeric H atom and the H atom localized at the glycosidic linkage point are, in the projection, in a parallel (synperiplanar) arrangement (Glcβ1,4Glc in Figure 23.5b; see also Figure 23.6).

With both β- and α-glycosidic residues R, C2 and R are each in the most stable, that is, antiperiplanar, conformation (at an angle of 180°) (Figure 23.7c or f). The two H atoms adjacent to the glycosidic O atom are in both cases approximately parallel (synperiplanar) (Figure 23.5b or a).

In the presence of an α-glycosidic linkage, the glycosidic O atom protrudes axially from C1 of the pyranose ring to the neighboring block. Again, a unique preferred direction results for the glycosidic residue R, due to the exo-anomeric effect. Likewise, in the presence of an α-glycosidic linkage of two monosaccharide units, the anomeric H atom and the H atom of the glycosidic linkage point are, in the projection, in parallel (synperiplanar) arrangement (Glcα1,4Glc in Figure 23.5a).

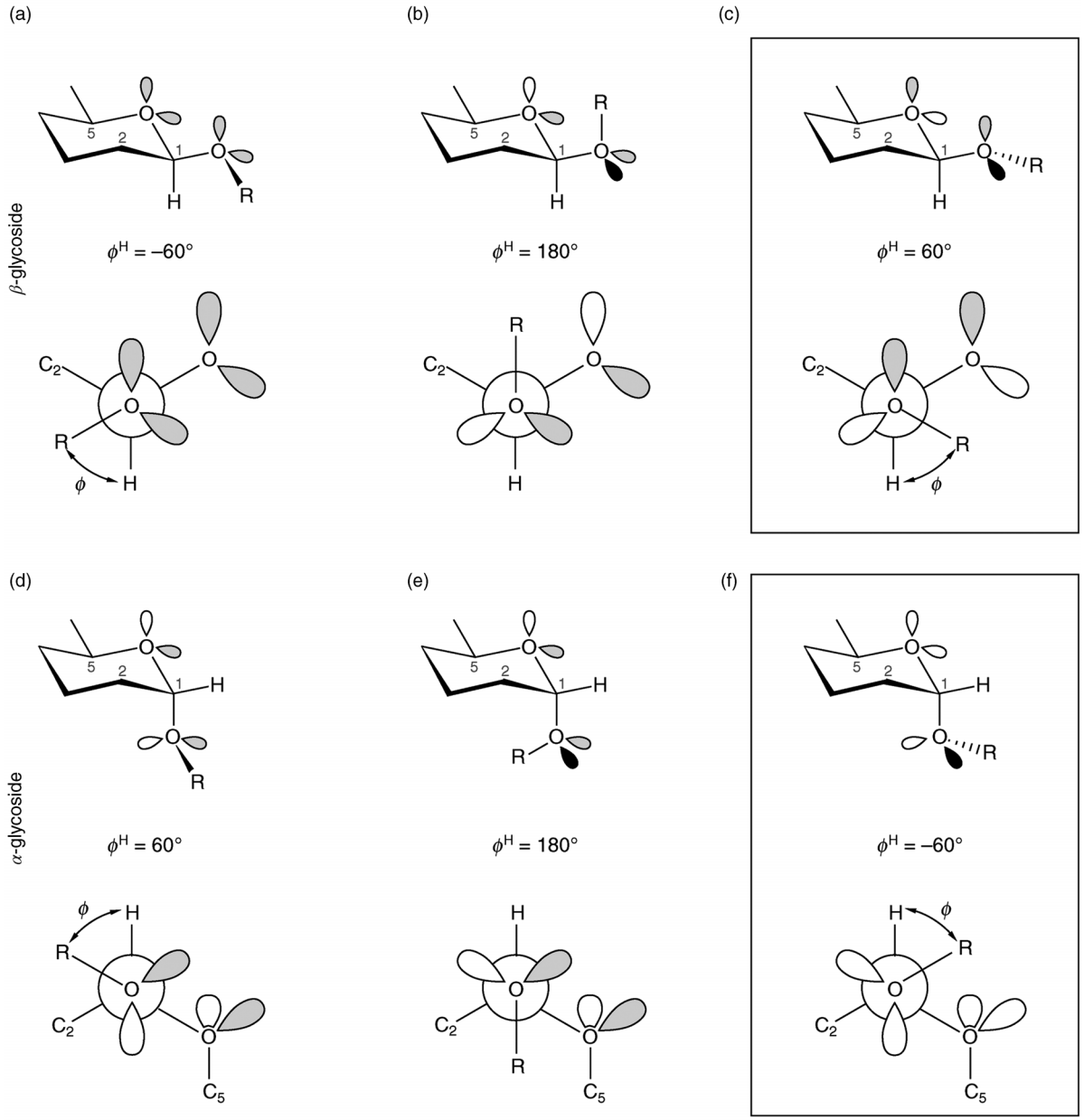


Figure 23.7 The exo-anomeric effect. The free pairs of electrons with parallel dipole moments are shaded. Important linkages that point towards the viewer are highlighted in bold. Important bindings that point away from the viewer are drawn with dashed lines.
β-Glycoside: (a) If the glycosidic residue R versus H1 of the pyranose ring takes an angle of $\Phi^H = -60°$ (a), then the dipole moments of the lone pairs of electrons localized at the two oxygen atoms each take the energetically unfavorable parallel arrangement, and the conformation is, so to speak, destabilized twice.(b) If the angle of R towards H1 is $\Phi^H = 180°$ (antiperiplanar arrangement of R and H1), then only one electron–dipole pair is placed in the energetically unfavorable parallel arrangement; consequently, (b) is energetically favored against (a). (c) If R and H1 take an angle of $\Phi^H = 60°$ (a simultaneous antiperiplanar arrangement of R and C2), then (as in case (b)) again only one electron–dipole pair is in the energetically unfavorable parallel arrangement; from a stereoelectronic point of view, (b) and (c) would be equal. In a stereochemical (spatial) view, however, it is apparent that the two major ligands C2 and R in (b) are spatially adjacent at an angle of 60°, while in (c) they face each other at an angle of 180° (antiperiplanar) and thus without steric interaction, which ultimately favors (c).
α-Glycoside: (d) If the glycosidic residue R relative to H1 of the pyranose ring is at an angle of $\Phi^H = 60°$, then at the two O-atoms one pair of electrons, each, is in the energetically unfavorable parallel arrangement of the dipole moments (dipole–dipole repulsion). (e) At an angle of $\Phi^H = 180°$, one electron–dipole pair is likewise in the energetically unfavorable parallel arrangement. The two α-glycosidic conformations (d) and (e) therefore appear, from a stereoelectronic point of view, energetically comparable. In addition, from a stereochemical point of viewpoint, (d) and (e) appear comparable to a first approximation, because the angle between R and C2 in both cases is 60°. On closer inspection, however, the glycosidic residue R in (e) is practically located under the monosaccharide unit (with correspondingly unfavorable stereochemical interaction), while in (d) it is located outside the monosaccharide unit, which results in an energetically more favorable situation. (f) If the glycosidic group R relative to H1 of the pyranose ring yields an angle of $\Phi^H = -60°$, then, however, none of the pairs of dipole moments at the two oxygen atoms is in parallel direction, which (from a stereoelectronic point of view) clearly favors (f), relative to (d) and (e). At the same time, the spatially demanding residues R and C2 are in an antiperiplanar configuration, which clearly favors (f) also from a stereochemical point of view.
Source: according to Lemieux, R.U. (1990) *Exploration with Sugars. How Sweet it Was*, American Chemical Society, Washington D.C.

23.2 Protein Glycosylation

With regard to the carbohydrate side chains, two different structural groups can be distinguished: *N*-glycans and *O*-glycans. The *N*-glycans are linked via *N*-acetyl-β-D-glucosamine (βGlcNAc) to the amide nitrogen of L-asparagine (Asn) (Figure 23.8a). For the initiation within a peptide sequence of an *N*-glycosylation, the amino acid sequence Asn-Xxx-Ser/Thr is required as a signal structure, where Xxx can be any amino acid except proline or aspartic acid. The *O*-glycans are linked via *N*-acetyl-α-D-galactosamine (αGalNAc) to the hydroxy group of L-serine (Ser) or L-threonine (Thr) (Figure 23.8b). According to current knowledge, there is no typical signal sequence to start *O*-glycosylation. In addition, there are special forms of glycosylation, which, however, are only mentioned briefly in this chapter.

Construction of the glycans in the cells takes place in the endoplasmic reticulum by enzymatic coupling of activated monosaccharide units to the respective precursor. In the course of the genetic evolution, a characteristic glycosylation pathway (a particular set of glycosyltransferases) has been selected in mammals, which causes a typical mammalian cell glycosylation.

Most glycoproteins have multiple *N*- and *O*-glycosylation sites, and each of these glycosylation sites usually carries several glycan structures. This means that a glycoprotein does not exist as a single structural form, but as a mixture of so-called glycoforms. These therefore have the same amino acid sequence (predetermined by genetic information), but differ in the number, position, and sequence of the attached glycans. Thus, different carbohydrate side chains generally occur on a specific asparagine – this is referred to as microheterogeneity.

Glycosylation is:

- Species-specific: for example, the glycosylation pattern of a human glycoprotein differs from the analogue glycoprotein of the hamster or mouse.
- Tissue-specific: the glycosylation pattern of the renal tissue, for example, differs from that of the ovary or of the connective tissue.
- Cell-type specific: the glycosylation pattern of a BHK cell differs from that of a CHO or C127 cell.
- Protein-specific: the glycosylation pattern of IFN-β is different, for example, from that of IL-2, or EPO, or AT-III, or t-PA (Table 23.1) – even if these glycoproteins are expressed in the same cell line (e.g., CHO cells) and cultured under identical conditions.

Thus, the study of the carbohydrate structures of glycoproteins for various research and development branches (such as glyco-biotechnology, molecular biology, clinical chemistry, medicine, and pharmacy) constitutes a major, challenging and, not least, charming area of responsibility.

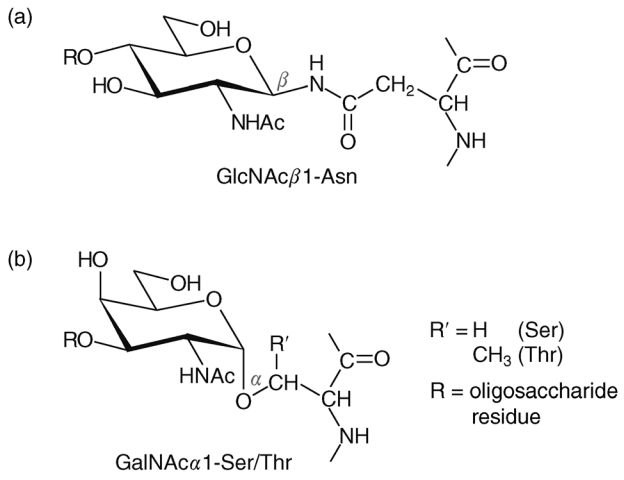


Figure 23.8 Linkage of glycans with the peptide backbone. (a) *N*-Glycans: they are linked to the amide nitrogen of L-asparagine (Asn) via *N*-acetyl-*β*-D-glucosamine (*β*GlcNAc). (b) *O*-Glycans: they are linked to the hydroxyl function of L-serine (Ser) or L-threonine (Thr) via *N*-acetyl-*α*-D-galactosamine (*α*GalNAc).

23.2.1 Structure of the *N*-Glycans

Basic types of *N*-glycans:
High-mannose type: the two α-Man residues of the pentasaccharide base structure are each occupied by mannose.

Complex type: the two α-Man residues of the pentasaccharide base structure are each occupied by GlcNAc.

Hybrid type: the two *α*-Man residues of the pentasaccharide base structure are occupied by mannose and GlcNAc (mixture of *high-mannose* and complex type).

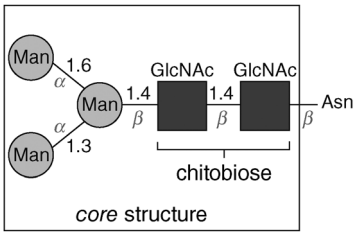


Figure 23.9 *N*-Glycan pentasaccharide Man_3-$GlcNAc_2$. The pentasaccharide basic structure consists of three components of D-mannose and two molecules of *N*-acetyl-D-glucosamine, which is N-glycosidically linked to the protein (each via amino acid asparagine) via the reducing end of the chitobiose *core* (GlcNAcß1,4GlcNAc) (see also Figure 23.8a).

The *N*-glycans of a glycoprotein are always *N*-glycosidically attached to the peptide backbone (via the sugar unit GlcNAc) via the amide group of the amino acid asparagine (Asn).

The glycans of glycoproteins of mammalian cells are constructed by glycosyltransferases following a uniform biochemical scheme, so that ultimately oligosaccharide side chains arise that are related to each other. All *N*-glycans share the basic pentasaccharide core structure Man_3-$GlcNAc_2$ (Figure 23.9). Fortunately (from the learner's perspective) there are only three basic types of *N*-glycans: the complex type, the high-mannose type (oligomannoside type), and the hybrid type (Figure 23.10). Within a glycoprotein, these branch types may be present simultaneously; however, under the influence of various factors (such as the current metabolic state of the cell), their ratio may be displaced in one or the other direction. This leads to the formation of the above-mentioned microheterogeneity. For example, earlier (so to speak "immature") stages of glycosylation carry more high-mannose types than later ("mature") stages of glycosylation, which identify themselves by their higher proportion of complex-type glycans.

The *N*-glycans of the high-mannose type structures possess only mannoses and two GlcNAc residues (Figure 23.10). The complex type *N*-glycan structure is shown in detail in Figure 23.11a, using a tetraantennary example.

The extraordinary variability even of this relatively uniform structural class results from the fact that these N-glycans can occur as follows:

- bi-, tri-, or tetraantennary (in exceptional cases also mono- and pentaantennary);
- with two different triantennary basic structures (2,4-branched or 2,6-branched);
- with Neu5Ac either 2,3- or 2,6-linked;
- with complete or incomplete sialylation;
- without or with 1–3 LacNAc repeat(s);
- with LacNAc repeats on different antennas or repetitive;
- with full structures on one or the other antenna;
- with defined changes in the individual antennas (e.g., by additional monosaccharide units and/or an unusual type of linkage or sulfation).

All these variants can occur both with and without "proximal fucose" and both with and without the so-called *bisecting* GlcNAc, which again quadruples the number of theoretically possible structures. (The "proximal fucose" is always bound to position 6 of the Asn-linked GlcNAc, the bisecting GlcNAc is always linked to position 4 of the β-linked core mannose.)

The hybrid type *N*-glycans virtually provide a mixture of a high-mannose type and a complex type sugar chain, where the complex type sugar chain, the antenna, is always attached to the α1,3-linked mannose of the pentasaccharide core.

The variability of the *N*-glycans is indeed quite diverse, their structural assignment (such as via mass spectrometry in conjunction with special mapping techniques) is, however, greatly simplified by their principally uniform basic structure.

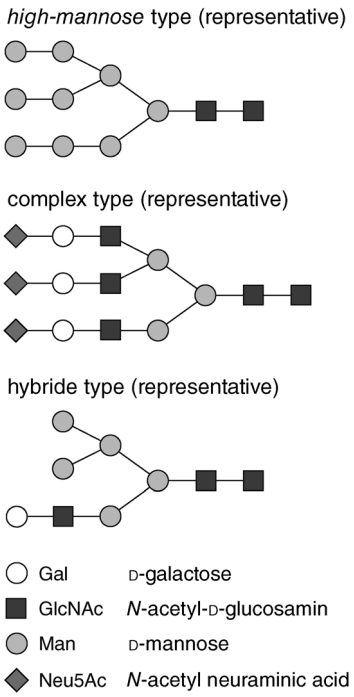


Figure 23.10 The three basic types of *N*-glycans.

23.2.2 Structure of the *O*-Glycans

The sequence and isomeric linkage of the monosaccharides in the *O*-glycans shows a greater variability than in the *N*-glycans. So far, at least eight different *O*-glycan types have been identified. The four most frequent basic types are shown in Figure 23.12.

The *O*-glycans of a glycoprotein are bound to the peptide backbone usually via *O*-glycosidic linkage of the sugar unit GalNAc to the OH group of the amino acid serine (Ser) or threonine (Thr). They are made in the first step by direct enzyme-catalyzed binding of GalNAc to serine or threonine, and then by sequential addition of other sugar units under the action of specific glycosyltransferases. Usually, one of the core structures shown in Figure 23.12 is found. The main one is core structure 1 (the so-called Thomsen–Friedenreich or T-antigen). At these core structures additional Gal or Neu5Ac residues may be attached, or the core structures may be extended with Galβ1,3GlcNAc (so-called *type I*) or Galβ1,4GlcNAc (*type II repeats*) where the repeats may be branched at C3 and C6 of the galactose unit. In addition, the *O*-glycans may still be sulfated, sialylated, and fucosylated.

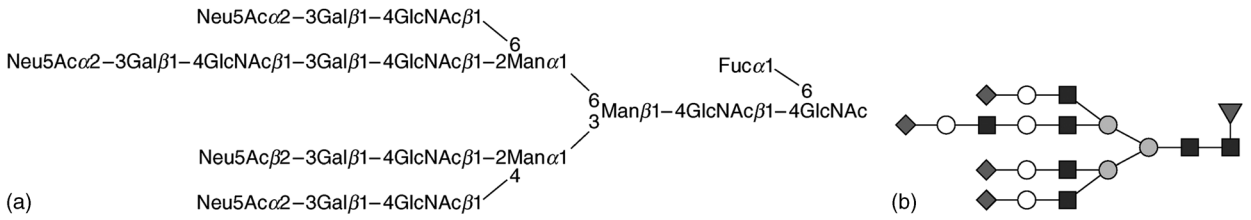


Figure 23.11 Example of a tetrasialo(α2,3) tetraantennary *N*-glycan with proximal fucose and one LacNAc-repeat (4N(2,3)-4A+1R). (a) Two-dimensional text nomenclature based on the rules of IUPAC (International Union of Pure and Applied Chemistry), modified according to the CFG (Consortium for Functional Glycomics). (b) Symbol nomenclature according to CFG. (see Figure 23.13).

O-Glycosylation may occur as a cluster, the mucin-type, which is often a component of extracellular matrices or secreted glycoproteins. Other *O*-glycosylations such as *O*-GlcNAc are found on many nucleoplasmic proteins, and *O*-mannose-linked glycosylation is found in some muscular and neural glycoproteins and in yeast. *O*-Fucose and *O*-glucose-linked glycosylation is found on many proteins similar to the epidermal growth factor.

So far, not much is known about the function of *O*-glycans. They contribute to the conformational stability of the glycoprotein and, together with the peptide backbone, they provide a characteristic structural motif that is important for structure recognition. Often they also provide the glycoprotein some protection against attack by proteases.

The characterization of *O*-linked oligosaccharides is substantially similar to that of the *N*-glycans. However, there is no easy distinction of the structural classes as in the *N*-glycans, and there is no universal enzyme by which the *O*-linked oligosaccharides may be cleaved.

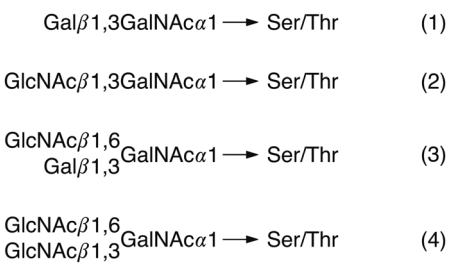


Figure 23.12 Core structures of *O*-glycans.

23.3 Analysis of Protein Glycosylation

The full characterization of a glycan consists of analysis of the primary structure with its branches, the configuration of the glycosidic bond per sugar unit (anomery), the linkage direction from sugar to sugar, and the assignment of isomers. However, many structural studies are facilitated by knowing the biochemical processes of glycosylation (e.g., equipment of a cell with specific glycosyltransferases).

In the following sections, erythropoietin (EPO) is often used as a model glycoprotein (Figure 23.13). Erythropoietin is a hormone (cytokine) that is formed in the kidney and controls the formation of red blood cells (erythropoiesis). Nowadays, it is produced recombinantly and used as a drug in the event of renal failure as well as after bone marrow transplantation and after aggressive cancer treatments. For this reason erythropoietin is, from an analytical point of view, probably the best-studied glycoprotein and, thereby, a natural candidate as a teaching example of glycoanalysis. Erythropoietin has three *N*-glycosylation sites (Asn24, Asn38, and Asn83) and one *O*-glycosylation site (Ser126), which have been extensively analyzed as part of product characterization for clinical approval.

First, we turn to the investigation of the intact glycoprotein.

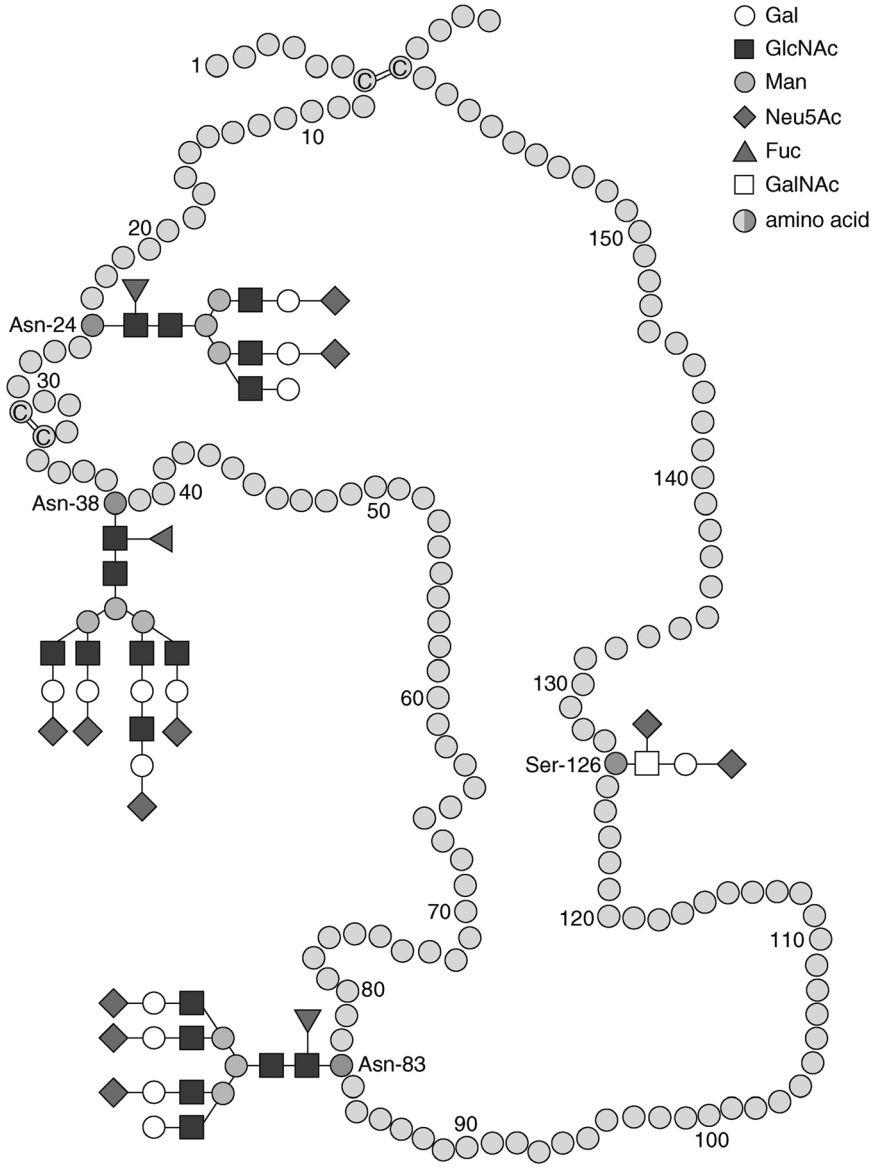


Figure 23.13 Schematic representation of erythropoietin. EPO is aglycoprotein having 165 amino acids with three *N*-glycosylation sites and one *O*-glycosylation site. The carbohydrate content is about 40% of the molecular weight. For recombinant expression of the cytokine in mammalian cells, rhEPO is secreted from the cells into the culture medium and can therefore relatively easily be isolated from the fermentation broth and purified.

23.3.1 Analysis on the Basis of the Intact Glycoprotein

Some glyco-analytical studies may already be performed on the basis of the intact (glyco) protein (Figure 23.14). Of course, you have first to determine whether the obtained molecule is in fact glycosylated. For this purpose, some of the methods described in the context of protein analysis are suitable, such as SDS-polyacrylamide gel electrophoresis (SDS-PAGE) (Section 11.3.9), isoelectric focusing (Section 11.3.12), or capillary electrophoresis (Chapter 12).

Electrophoretic Analyses The *N*-glycans of glycoproteins can be cleaved from the peptide backbone, for example, using the enzyme PNGase F (peptide *N*-glycosidase F from *Flavobacterium meningosepticum*) (Section 23.3.3). Because the loss of one or more sugar chains causes a loss of molecular weight, the behavior of the glycoprotein in SDS-PAGE will change – its apparent molecular weight will decrease. This is seen particularly well with EPO as an example (Figure 23.15).

SDS Page, Section 11.3.9

Isoelectric Focusing, Section 11.3.12, 12.4.7

Another possible means of investigation is via isoelectric focusing (IEF). In IEF, proteins and glycoproteins are separated in a stable pH gradient according to their isoelectric point. The

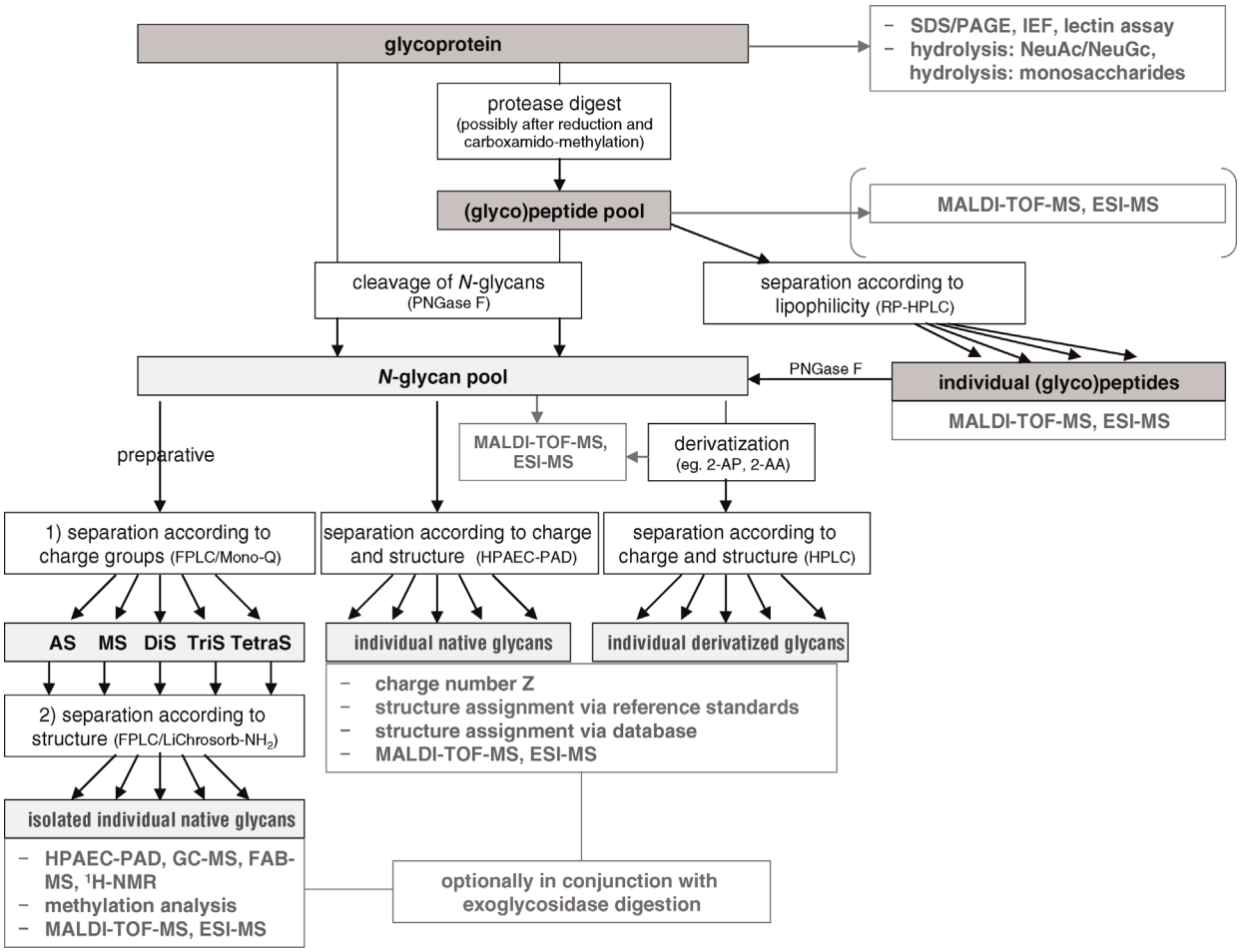


Figure 23.14 Methodological approaches to the analysis of protein glycosylation.

glycoforms of a glycoprotein differ (due to the identical peptide backbone) only by the type and weighting of the sugar chains at the individual glycosylation sites – and thus by the number of negatively charged Neu5Ac residues (or sulfate residues, if available). Thus, IEF of the intact glycoprotein allows an important statement to be made with respect to the sialylation state and the percentage share of the glycoforms, each differing by one charge, which again is particularly evident from EPO as an example (Figure 23.16).

The glycoforms of a glycoprotein may also be separated and quantified via capillary electrophoresis (Chapter 12). Thus, the international EPO reference preparation, separated

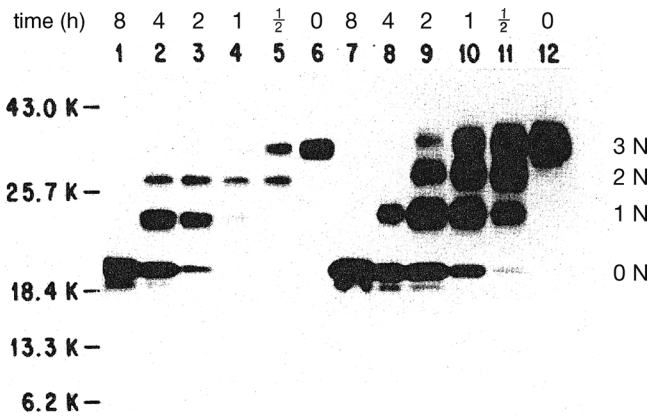


Figure 23.15 Time course of treatment of urinary EPO (lanes 1–6) and recombinant EPO (lanes 7–12) with PNGase F; analysis via 12.5% SDS-PAGE and Western blot. The cleavage of each of the three *N*-glycosidically linked sugar chains (N) shifts the bands to smaller molecular weight (3N → 2N → 1N → 0N). After cleavage of the *N*-glycans, the residual molecule migrates as two bands. The larger (upper) band corresponds to the *N*-deglycosylated peptide backbone, which is *O*-glycosylated at Ser-126, while the smaller (lower) band represents the completely sugar-free peptide backbone. The two bands show that the EPO peptide backbone is *O*-glycosylated only at about 70%. This fact plays an important role in batch consistency control of rhEPO, but it need not concern us in this chapter. Source: according to Egrie, J.C. *et al.* (1986) *Immunobiology*, **172**, 213–224. With permission, Copyright © 1986 Gustav Fischer Verlag · Stuttgart · New York. Published by Elsevier GmbH.

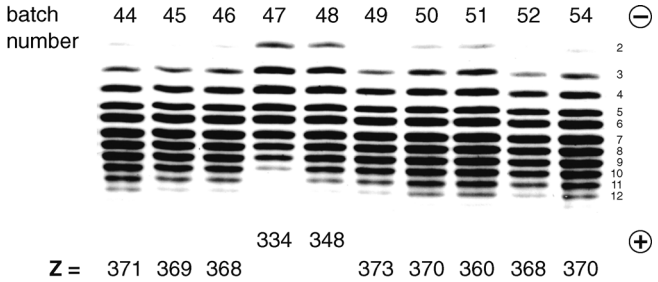


Figure 23.16 Isoelectric focusing pattern of recombinant human erythropoietin from C127 mouse fibroblast cells. Ten successive rhEPO production batches were analyzed in the context of quality control analytics by IEF and HPAEC-PAD (cf. Section 23.3.4). The focusing patterns of these samples showed about ten glycoforms of defined negative charge, possibly with a band focus in isoform 7. It was found in the material of batch 47 that bands 12 and 11 were missing (the glycoforms with the highest negative charge) and instead a new glycoform (band 2) was detected that had ten negative charges less than band 12. The band pattern of batch 48 also exhibited a loss of negative charge, although to a lesser extent than in batch 47, while the glycoform pattern of the remaining batches were roughly comparable. This loss of negative charge in lots 47 and 48 was even more apparent from the mapping profiles of the isolated rhEPO *N*-glycan pool (Figure 23.24) and could be quantified via the hypothetical charge number *Z* (*Z-number*) that is presented in Section 22.3.4. While for the eight normal batches an average charge number of $Z = 369 \pm 4.1$ (CV = 1.1 %) was found, the charge numbers for batches 47 and 48 were only 334 and 348, respectively, which especially for batch 47 pointed towards a less favorable glycosylation (in terms of biological activity). Source: according to Hermentin, P. and Witzel, R. (1999) *Pharm. Pharmacol. Commun.*, **5** 33–43. With permission, © 1999 Royal Pharmaceutical Society of Great Britain

according to the standard conditions described in the European Pharmacopoeia, yields eight charge-separated isoforms that can be quantified via their peak areas, thus enabling a statement with respect to the sialylation status (Figure 23.17).

Analysis using Lectins The oligosaccharide chains of glycoproteins may, to a certain extent, be characterized by reaction with lectins. Lectins are (glyco)proteins of plant origin that specifically recognize and bind to carbohydrates and thus allow a selective glyco-analysis. If you highlight a lectin of known carbohydrate specificity with a suitable marker (e.g., digoxigenin), you can recognize the carbohydrate structures detected by the lectin in a sandwich assay. Today a series of lectins with sufficiently characterized specificity is available for such a purpose. An appropriate experimental setup, similar to the one shown in Figure 23.18, allows us to recognize the terminal sugar residues of the glycans specifically and to visualize them in a color reaction. Due to the microheterogeneity of glycosylation mentioned in Section 23.2, this method is, however, limited in its meaningfulness.

The lectin method consists of the following steps:

1. separation of the protein or protein mixture via gel electrophoresis;
2. blotting of the proteins onto a PVDF membrane;

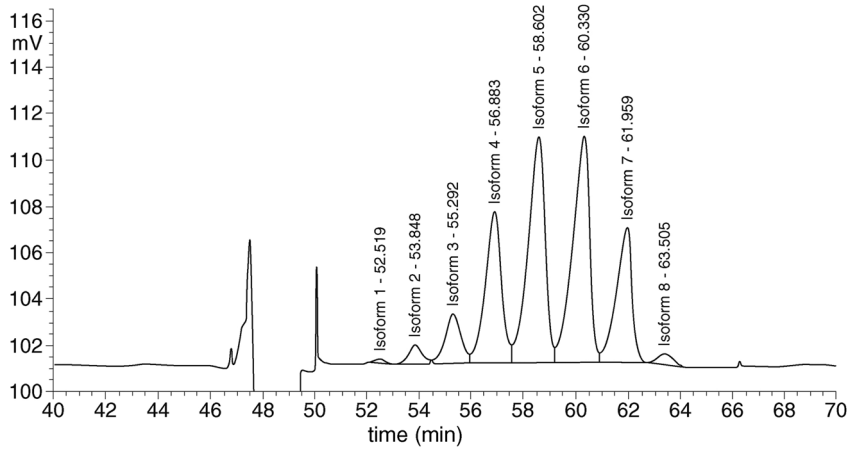


Figure 23.17 Capillary electrophoretic separation of the EPO (CHO) reference standard at the standard conditions prescribed in the European Pharmacopoeia. Source: according to: Behr-Gross, M.E. *et al*. (2007) *Pharmeuropa Bio*, (1), 49–66. © EDQM/European Directorate for the Quality of Medicines & Health-Care.

3. incubation with a lectin conjugated with digoxigenin;
4. detection of the digoxigenin in an immune reaction with an anti-digoxigenin antibody (e.g., polyclonal from rabbit or monoclonal from mouse) conjugated with an enzyme (such as alkaline phosphatase, AP);
5. sugar-specific staining of the protein band via the enzyme immobilized in the sandwich.

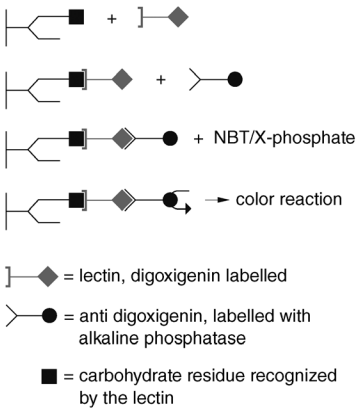


Figure 23.18 Sandwich assay with lectins to characterize the glycosylation of a glycoprotein. The lectin–digoxigenin conjugate binds selectively to the sugar residues of the glycoprotein recognized by the lectin. The thus labeled glycoprotein is detected with an anti-digoxigenin antibody that is conjugate with alkaline phosphatase (AP). The addition of NBT results in a specific staining of the glycoprotein.

Table 23.3 shows the results of lectin analyses of four different glycoproteins: The lectin GNA (with specificity for α-linked mannose) stains only carboxypeptidase Y, which therefore (in contrast to transferrin, fetuin and asialofetuin) obviously carries high-mannose type structures. The lectin SNA, which recognizes α-2,6-linked Neu5Ac residues, reacts with both transferrin and fetuin, and these two glycoproteins therefore have glycans with the corresponding terminal structures. In contrast thereto, the lectin MAA, which is specific for α-2,3-linked Neu5Ac residues, proves this structure only for fetuin. Both lectins (SNA and MAA) do not react with desialylated fetuin (asialofetuin), which, therefore, acts as a negative control and thus excludes a false-positive result. The lectin DSA (with specificity for Galβ1,4 and GlcNAcβ1,4) reacts with transferrin, fetuin, and asialofetuin and thus discloses for these glycoproteins the presence of complex type *N*-glycans. *O*-Glycans can be detected after removal of sialic acid residues: The peanut agglutinin PNA recognizes the non-sialylated Galβ1,3GalNAcα1-structures of the *O*-glycans and thus shows that fetuin – in addition to the *N*-glycans demonstrated by the reaction with DSA – also has *O*-glycosidically linked sugar chains.

Analysis of the Monosaccharide Components Quantitative methods for determination of the monosaccharide components analyze the relative proportions of neutral sugar units of a glycoprotein. This allows to some extent the determination of the structure of the glycans present.

To determine the monosaccharide components, the sample should be free of salts and additives (biopharmaceutical products sometimes contain mono- or disaccharides as carriers). Low molecular weight components may be purified from the sample by dialysis against water or volatile buffer (Section 1.6) or by gel chromatography and elution with water or volatile buffer (Section 10.4) or solid phase extraction on a C18 or C8 matrix (Section 24.2.2).

The actual analysis is performed either with the intact glycoprotein (without prior degradation) or after acid hydrolysis of the oligosaccharide chains, which gives qualitative and quantitative information on the monosaccharide species present.

Analysis of the Neutral Monosaccharide Components The common method used to quantify the neutral sugars in glycoproteins is based on the formation of a dye that is produced by heating the glycoprotein in the presence of aqueous phenol in concentrated sulfuric acid. In many cases, the heat required for the reaction is generated by the addition of concentrated sulfuric acid to the glycoprotein–phenol mixture. A fast and efficient mixing of the solutions is critical for

Table 23.3 Lectin analysis of four different glycoproteins.[a)]

	Lectin:				
	GNA	SNA	MAA	DSA	PNA
Glycoprotein					
Carboxypeptidase Y	+	–	–	–	–
Transferrin	–	+	–	(+)	–
Fetuin	–	+	+	+	–
Asialofetuin	–	–	–	+	+
Specificity					
	Manα	Neu5Ac	Neu5Ac	Gal	Galβ1, 3GalNAc (asialo-*O*-glycan)
		α2,6	α2,3	β1,4	
				GlcNAc	
				β1,4	

a) + positive reaction; – no reaction.

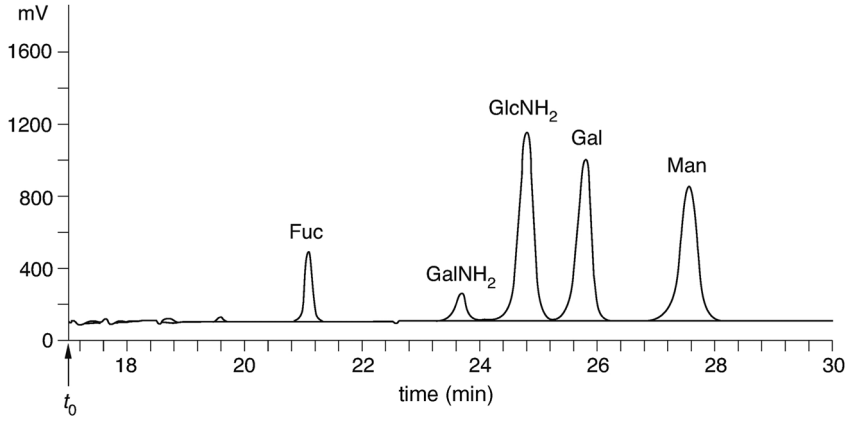


Figure 23.19 HPAEC-PAD profile of the major neutral monosaccharide components of mammalian cell glycoproteins: L-fucose (Fuc), D-galactosamine ($GalNH_2$), D-glucosamine ($GlcNH_2$), D-galactose (Gal), and D-mannose (Man). Of these sugars, fucose elutes very early because – in comparison to the other hexoses – at C6 it has an H-atom in place of an OH-group, which means that the interaction with the anion-exchange matrix is substantially reduced. The monodeoxy-hexoses and pentoses elute comparatively earlier than L-fucose (6-deoxy-L-galactose) as they carry one OH-group less than hexoses (t_0: sample injection).

consistent analysis. Quantitative results are obtained by the parallel analysis of standards used to generate a standard curve by which the dye formation can be quantified.

Somewhat more complex but more meaningful is detection by high performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD) of the monosaccharide components liberated from the glycoprotein (Figure 23.19). In comparison to HPLC or gas chromatography, and also compared to capillary electrophoresis, HPAEC-PAD has the advantage that it does not require derivatization of the released monosaccharides, as these can directly be detected and measured quite specifically by means of an amperometric detector.

This anion-exchange chromatography via pellicular matrices carrying aminopropyl ligands (Carbo-Pac HPLC column) separates the monosaccharides liberated from the glycoprotein in alkaline conditions, where the sugar chains occur as negatively charged alcoholates and therefore interact with the anion-exchange matrix. The analysis of the carbohydrate structures is performed in a very selective and sensitive manner – without derivatization – by pulsed amperometric detection (PAD) or pulsed electrochemical detection (PED) at a gold electrode. Thereby, the redox potentials of the detector have to be adjusted to the redox potentials typical for the oxidation of sugars.

As polyhydroxy compounds the neutral monosaccharides are weak acids with ionization constants of 10^{-12} to 10^{-14}. Monosaccharides can therefore be separated via such an anion-exchange column at pH >13, whereby the interaction with the matrix is determined by the acidity of the respective sugar molecule. The acidity of the individual OH groups decreases in the order OH1 > OH2 ≥ OH6 > OH3 > OH4.

HPAEC-PAD does not require any derivatization of the monosaccharides and thus is simpler and superior to other methods of monosaccharide analysis (Figure 23.19).

With mammalian cell glycoproteins, acid hydrolysis in 2 N trifluoroacetic acid (4 h, 100 °C) with subsequent detection and quantification of the monosaccharides via HPAEC-PAD has proven optimal. Trifluoroacetic acid (TFA) has, compared to HCl or H_2SO_4, the advantage that it is volatile and removed during lyophilization or in a SpeedVac. At the given acid hydrolysis conditions, *N*-acetyl sugars (as GlcNAc or GalNAc) will be *N*-deacetylated and detected as the corresponding amino sugars ($GlcNH_2$ or $GalNH_2$). Since *N*-acetylneuraminic acid decomposes under these hydrolysis conditions, Neu5Ac must be released in a special and milder hydrolysis or enzymatically (e.g., by means of neuraminidase (sialidase)), and determined separately (see below).

As is obvious from the *N*-glycan structures (Figures 23.9–23.11), defined monosaccharide molar ratios may be concluded from the regular biantennary, triantennary, and tetraantennary *N*-glycans with complete sialylation. The same applies for the *high-mannose* type *N*-glycans (Table 23.4). In some cases the molar GlcNAc/mannose ratio may even enable us to conclude the actual structure type: A GlcNAc/Man < 0.5 ratio generally points towards the presence of *high-mannose* type *N*-glycans, whereas at a ratio between 1 and 2 complex type *N*-glycans may predominantly (or exclusively) be expected.

For rhEPO the quotient GlcNAc/Man = 2.0 indicates the presence of complex type *N*-glycans, where a high proportion of tetraantennary *N*-glycans may be expected. The finding of mannose (9 mole $mole^{-1}$) points towards the presence of three glycosylation sites, as complex type *N*-glycan structures always carry three mannose units per *N*-glycan. Likewise, the GlcNAc content (18 mole $mole^{-1}$) and Gal content (14 mole $mole^{-1}$) indicate the presence of three tetraantennary *N*-glycans.

Table 23.4 Molar ratios of monosaccharide components of *N*-glycans.

		Fuc	GalNAc	GlcNAc	Gal	Man	Neu5Ac	GlcNAc/Man	Neu5Ac/Gal	Man/Fuc
Complex type	Tetraantennary	0–1	—	6	4	3	4	2.00	1.00	
	Triantennary	0–1	—	5	3	3	3	1.67	1.00	
	Biantennary	0–1	—	4	2	3	2	1.33	1.00	
High-mannose type	Man5	—	—	2		5		0.40		
	Man6	—	—	2		6		0.33		
	Man7	—	—	2		7		0.29		
	Man8	—	—	2		8		0.25		
	Man9	—		2		9		0.22		
Erythropoietin		3	0.8	18	14	9	11	2.0	0.8	3.0

The fact that approximately 1 mole per mole of GalNAc is detected for rhEPO indicates the presence of a single *O*-linked glycosylation site. Accordingly, 1 mole per mole of galactose should be attributed to the *O*-glycan (Section 23.2.2), which means that 13 mole per mole of galactose should still remain for the three tetraantennary *N*-glycans.

The detection of GalNAc generally indicates the presence of *O*-glycans, because GalNAc is only in exceptional cases part of *N*-glycans. In the case of positive GalNAc findings, the presence of an *O*-glycan should be confirmed using another method. On the other hand, a negative GalNAc result always proves the lack of any *O*-glycosylation.

The data of a monosaccharide components analysis should not be overestimated! In general, the situation will be much more complicated, and usually no information on the nature of the oligosaccharide side chains is obtained from monosaccharide composition analysis.

Analysis of Sialic Acid Assessment of the sialylation status is of great interest, especially with biopharmaceutical glycoproteins such as EPO, because this determines the half-life of the glycoprotein in the blood circulation and thus its biological activity. Incompletely sialylated glycoproteins are recognized by receptors in the liver and removed from the blood circulation (*clearance*), and their biological activity is thus reduced. The production of therapeutic glycoproteins therefore demands confirmation of the intact and consistent glycosylation and sialylation from batch to batch for pharmacological safety reasons.

In the thiobarbiturate method according to Warren (the so-called Warren test) the sialic acid residues are released from the glycoprotein by acid hydrolysis, oxidized by periodate, and stained by the addition of thiobarbituric acid. The color may likewise be generated by the reaction of neuraminic acid with resorcinol. The resulting colored complex is extracted from the aqueous medium with cyclohexanone or *n*-butanol/butyl acetate (1 : 5). In both cases the result is quantified via sialic acid standards carried in the reaction. The absorption of the respective color complex measured at 549 and 580 nm is, over a certain concentration range, proportional to the sialic acid content of the sample.

Nowadays, these two staining methods have been largely replaced by the simpler and clearer analysis via HPAEC-PAD (see above). The bound sialic acid residues are liberated from the glycoprotein by mild acid hydrolysis (e.g., in 0.1 N H_2SO_4 or 2 N acetic acid for 0.5–1 h at 80 °C), or by incubation with the enzyme neuraminidase. The actual analysis is then performed via HPAEC-PAD, which not only enables the exact quantification using a calibration curve but also a simple differentiation between *N*-acetylneuraminic acid (Neu5Ac) and *N*-glycolylneuraminic acid (Neu5Gc) (structural formulas Figure 23.3). Compared to the *N*-acetyl group, the *N*-glycol group carries an additional OH group that is acidified by the adjacent CO group. This causes for Neu5Gc an enhanced interaction with the anion-exchange matrix and thus a greatly increased retention time.

Neu5Gc does not occur in natural human glycoproteins, but does, however, for example, in bovine glycoproteins or monoclonal antibodies derived from mouse cell lines. The NeuGc content of human glycoproteins recombinantly expressed in animal cells is very low (in the case of rhEPO (CHO) less than 3%); nevertheless, for biopharmaceutical glycoproteins this provides an important quality parameter that needs to be controlled from batch to batch.

Under the alkaline conditions of routine HPAEC-PAD, *O*-acetylated neuraminic acid residues lose their *O*-acetyl groups (ester hydrolysis). Its detection therefore requires neutral or weakly acidic eluents or (better) a mass spectrometric analysis of the corresponding glycopeptides (an example is given below in Figure 23.22c).

23.3.2 Mass Spectrometric Analysis on the Basis of Glycopeptides

Direct mass spectrometric characterization of a glycoprotein is complicated by the microheterogeneity of glycosylation. Multiply and heterogeneously glycosylated proteins (such as EPO) lead to a broad, unresolved mass peak. However, on the basis of glycopeptide fragments, mass spectrometric methods (MALDI-MS Section 15.1, ESI-MS Section 15.2) in combination with HPLC are essential. They enable quick insight into the glycan composition per glycosylation site (microheterogeneity) and provide important sequence information via fragmentation experiments (MS/MS) (Section 23.3.4).

Mass Spectrometry, Chapter 15

For this purpose, the glycoprotein to be analyzed is digested with an appropriate protease (e.g., trypsin, chymotrypsin, endoproteinase LysC, endoproteinase GluC; cf. Section 9.5.1). Direct characterization of the (glyco)peptide pool via MS, however, is limited by a masking effect (*ion suppression*) the peptide signals exert on the glycopeptide signals. Therefore, the (glyco)peptide pool is usually separated via RP-HPLC and the individual peaks are analyzed via mass spectrometry, either online (LC/MS) or, after fractionation, offline (MALDI-MS, ESI-MS). MS analysis of glycopeptides plays an important role, particularly in the characterization of *O*-glycans, since the free *O*-glycans are not readily accessible as such (Section 23.2.2).

Trypsin cleaves peptide bonds of polypeptides C-terminal from lysine (K) and arginine (R). If, for example, EPO is digested with trypsin, two glycosylation sites (Asn24 and Asn38) remain on one and the same fragment (T5, amino acid residues 21–45).

Endoproteinase GluC cleaves peptide bonds each C-terminal from glutamate (E) and occasionally aspartate (D). If you digest EPO with endo-GluC, the three *N*-glycosylation sites and the *O*-glycosylation site will be located on four different fragments (E5, E6, E10, and E12, Figure 23.20). The resulting 13 endo-GluC fragments can be separated via RP-HPLC (Figure 23.21) and allow the determination of the site-specific glycosylation (glycosylation per glycosylation site).

Mass spectrometric analysis of the (glyco)peptides separated via RP-HPLC (see example of EPO, Figure 23.21) allows identification of the various glycosylation sites: The mere peptides can be assigned by comparing the experimental masses to the masses calculated from the respective amino acid sequence; the non-attributable peptide masses refer to an existing glycosylation (see endo-GluC digest of EPO, Table 23.5).

The mass spectra of the glycopeptides particularly enable us to deduce the glycosylation per glycosylation site, which allows, via the molecular masses detected, assignment of the site-directed microheterogeneity (see example of EPO, Figure 23.22). Thus, the mass analysis of

$_{1}$APPRLI**C**D$_{8}$ **(E1)** SR$_{10}$VLE$_{13}$ **(E2)** RYLLE$_{18}$ **(E3)** AK$_{20}$E$_{21}$ **(E4)**
AE**N**$_{24}$ITTG**C**A$_{30}$EH**C**SLNE$_{37}$ **(E5)** **N**$_{38}$IT$_{40}$VPD$_{43}$ **(E6)** TKVNFYA$_{50}$WKRME$_{55}$ **(E7)**
VGQQA$_{60}$VE$_{62}$ **(E8)** VWQGLALL$_{70}$SE$_{72}$ **(E9)** AVLRGQAL$_{80}$LV**N**$_{83}$SSQPWEP$_{90}$LQLHVD$_{96}$ **(E10)**
KAVS$_{100}$GLRSLTT$_{107}$ **(E11)** LLR$_{110}$ALGAQKEAIS$_{120}$PPDAAS$_{126}$AAPL$_{130}$RTITAD$_{136}$ **(E12)**
TFRK$_{140}$LFRVYSNFLR$_{150}$GKLKLYTGE$_{159}$ **(E13)** A$_{160}$**C**RTGD$_{165}$ **(E1)**

Figure 23.20 Amino acid sequence of EPO with GluC cleavage products E1–E13 (*N*-glycosylation at Asn24, Asn38, and Asn83; *O*-glycosylation at Ser126; disulfide bond in the two peptides E1 [(1–8)SS(160–165)] and E5).

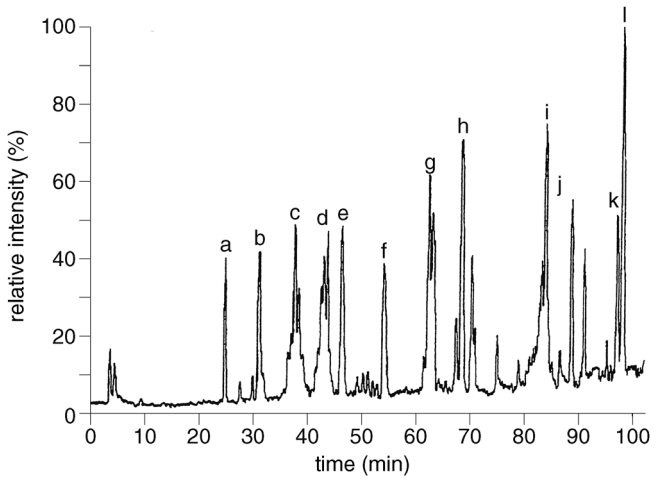


Figure 23.21 Total ion current (negative ion mode) of the (glyco)peptides of an endo-GluC digest of EPO (CHO) separated by means of RP-HPLC/ESI-MS. The peaks are detected as single or double charged species. Peaks c, d, g, and i correspond to the glycopeptide fragments of the four glycosylation sites; the observed peak broadening or peak splitting results from the microheterogeneity of glycosylation. Source: according to: Kawasaki, N. *et al.* (2000) *Anal. Biochem.*, **285**, 82–91. With permission,

Table 23.5 Theoretical and experimental masses of EPO GluC peptides; fragments E5* (Asn24), E6* (Asn38), E10* (Asn83) and E12* (Ser126) are glycosylated. Source: according to Kawasaki, N. *et al.* (2000) *Anal. Biochem.*, **285**, 82–91.

Peak in Figure 23.21	Glu-C peptide number	Amino acid residue (cf. Figure 23.20)	Average theoretical mass	*m/z*	
				M^{1-}	M^{2-}
A	E8	56–62	729.4	728.4	
B	E2	9–13	602.4	601.5	
C	E6*	38–43			
D	E5*	22–37			
E	E1	(1–8)S-S(160–165)	1502.7	1502.1	750.6
F	E3	14–18	692.1	691.5	
G	E12*	108–136			
H	E7	44–55	1571.8	1571.1	
I	E10*	73–96			
J	E9	63–72	1114.6	1113.9	
K	E13	137–159	2835.6		1417.0
L	E11	97–107	2211.3	2210.6	

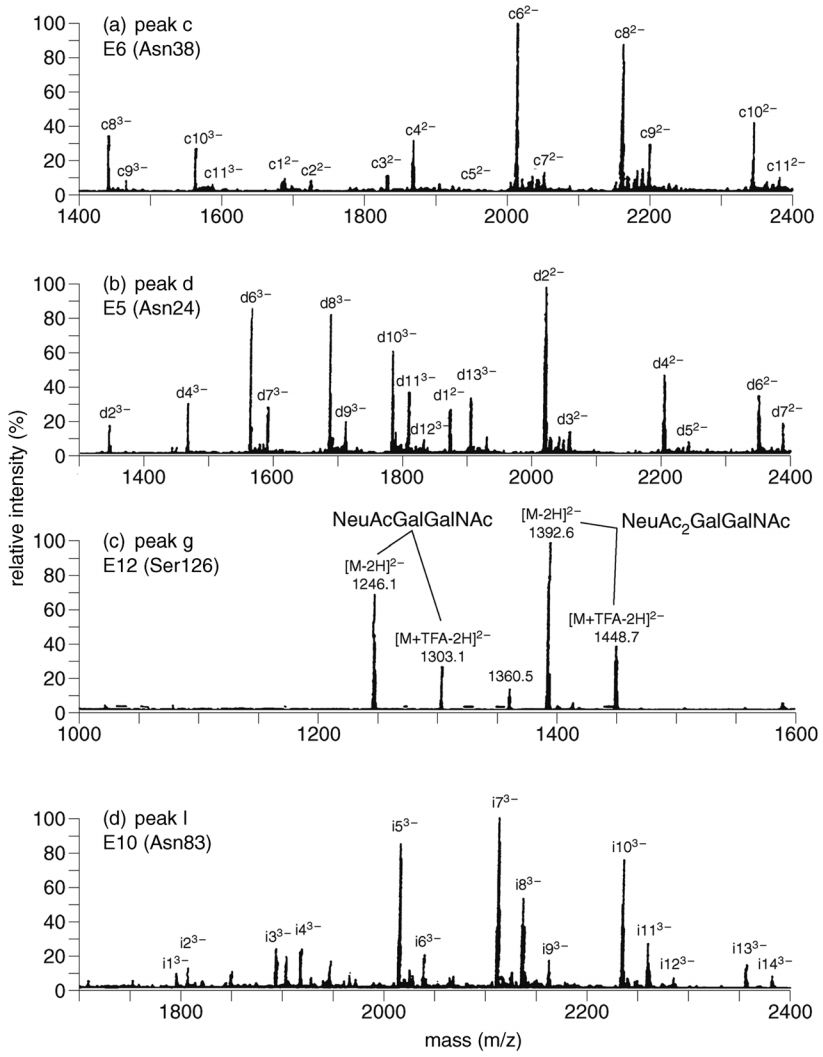


Figure 23.22 ESI-MS mass spectra of the four glycopeptides of GluC digestion of EPO (CHO) from Figure 23.21: peak c (a), peak d (b), peak g (c), and peak i (d). The assigned *N*-glycan structures are shown in Table 23.6. Source: according to: Kawasaki, N. *et al*. (2000) *Anal. Biochem.*, **285**, 82–91. With permission, Copyright © 2000 Academic Press.

Table 23.6 Theoretical and experimental masses of EPO GluC glycopeptides; E5 (Asn24), E6 (Asn38) and E10 (Asn83) (peaks c, d and i of Figure 23.21). Source: according to Kawasaki, N. *et al.* (2000) *Anal. Biochem.*, 285, 82–91.

Ion in Figure 23.21	Amino acid residue (cf. Table 23.5)	*N*-Glycan structure (each with *core*-fucose)	Average theoretical mass	*m/z*	
				M^{2-}	M^{3-}
c	E6 (38–43)	Asn_{38}			
c1		$BiLac_1NA_2$, $TriNA_2$	3375.2	1686.8	
c3		$TriNA_3$	3666.5	1831.8	
c4		$BiLac_2NA_2$, $TriLac_1NA_2$, $TetraNA_2$	3740.6	1869.1	
c6		$TriLac_1NA_3$, $TetraNA_3$	4031.8	2014.5	
c8		$TetraNA_4$	4323.1	2160.2	
c10		$TetraLac_1NA_4$	4688.4	2343.2	1561.8
d	E5 (22–37)	Asn_{24}			
d1		$BiNA_1$	3750.7	1875.1	
d2		$BiNA_2$	4042.0	2020.3	1346.2
d4		$BiLac_1NA_2$, $TriNA_2$	4407.3	2202.7	1468.2
d6		$TriNA_3$	4698.6	2348.7	1565.6
d7		$BiLac_2NA_2$, $TriLac_1NA_2$, $TetraNA_2$	4772.6	2385.7	1590.7
d8		$TriLac_1NA_3$, $TetraNA_3$	5063.9		1687.3
d10		$TetraNA_4$	5355.2		1784.5
d13		$TetraLac_1NA_4$	5720.5		1905.7
i	E10 (73–96)	Asn_{83}			
i1		$BiLac_1NA_2$, $TriNA_2$	5388.5		1795.1
i3		$TriNA_3$	5679.8		1892.2
i4		$BiLac_2NA_2$, $TriLac_1NA_2$, $TetraNA_2$	5753.9		1917.6
i5		$TriLac_1NA_3$, $TetraNA_3$	6045.1		2014.2
i7		$TetraNA_4$	6336.4		2111.1
i10		$TetraLac_1NA_4$	6701.7		2233.4
i13		$TetraLac_2NA_4$	7067.1		2355.8

HPLC peak c of Figure 23.21 (GluC cleavage product E6 of EPO with glycosylation at Asn38) reveals 11, peak d (E5 with Asn24) 13, and peak i (E10 with Asn83) 14 different glycostructures, the major masses of which are shown in Table 23.6. Thereby, the individual ion peaks may cover different glycan isomers that cannot be identified in this glycopeptide analysis, but may only be assigned on the basis of the free *N*-glycans (cf. Section 23.3.5). Only for peak g (E12 with *O*-glycosylation at Ser126) do the *m/z* values refer to the structures NeuAc-Gal-GalNAc and $NeuAc_2$-Gal-GalNAc (Figure 23.22c).

23.3.3 Release and Isolation of the *N*-Glycan Pool

The procedure for glycan release depends on the glycoprotein to be analyzed and the degree of the desired information. Depending on the target, different enzymatic as well as chemical methods may be applied.

Thus, the asparagine-linked sugar chains of glycoproteins may be released by glycopeptidases (PNGase F, PNGase A) or endoglycosidases (Endo-H, Endo-F) or by heating in anhydrous hydrazine for several hours.

The *N*-glycans separated from the peptide backbone may be analyzed in chromatographic or electrophoretic manner or by MS or a combination of these methods.

Note: For the release of *O*-glycans, there is currently only a single enzyme, *O*-glycanase from *Diplococcus pneumoniae* – but this enzyme cleaves, because of its high substrate specificity, only Galβ1,3GalNAcα-Ser/Thr. In addition, an ideal chemical release process does not exist. However, *O*-glycans can usually be released by reductive alkali-catalyzed β-elimination (alkali borohydride reaction); the free glycans are reduced *in situ* to avoid the formation of degradation products. But this reaction is not specific so that in parallel 10–20% of the available *N*-glycans will be released. Thereby, the free glycans lose their reactive reducing ends, so that these are no longer available for the easy introduction of a UV or fluorescent label. However, the reduced *O*-glycans may be easily identified by mass spectrometry, so that their analysis today is usually carried out by MS.

Enzymatic Release by PNGase F The enzyme peptide-N^4-(*N*-acetyl-β-glucosaminyl) asparagine amidase (PNGase F from *Flavobacterium meningosepticum*) practically cleaves all *N*-glycosidically linked sugar chains (except those bound to the amino or carboxy terminus of a polypeptide). (Also excluded are those *N*-glycans that at the proximal GlcNAc carry α1,3 (instead of α1,6)-linked fucose, as found in plant and insect cells but not in animal cells; such *N*-glycans are released by PNGase A). However, the enzymatic process sometimes requires optimized reaction conditions – such as a prepend proteolytic digestion of the glycoprotein (e.g., with trypsin or chymotrypsin). The peptide fragments thus formed are conformationally more flexible, and thus enable easier access of the enzyme to the individual glycosylation sites. The addition of an appropriate detergent (e.g., Triton X-100, Tween 20, CHAPS, or sodium dodecyl sulfate) facilitates the cleavage of the sugars (the detergent will unfold the protein and thus likewise facilitate the access of the enzyme to the glycosylation sites). PNGase F cleaves the *N*-glycosidic bond between the "proximal GlcNAc" (the anchor sugar) and the amino acid asparagine (Figure 23.23). While here the *N*-glycans are released in unaltered form, the anchor molecule asparagine (Asn) will be modified to aspartic acid (Asp).

The removal of the sugar residues by PNGase F will usually cause a decrease in molecular weight, which is evident in SDS-PAGE (see example of EPO, Figure 23.15). The released sugar chains can be extracted with (freezer-cooled) ethanol (<10 °C), desalted in a further purification step (e.g., via a Hypercarb™ cartridge), and then subjected to the actual analysis.

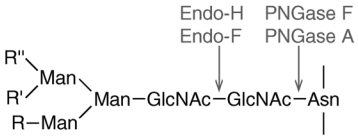


Figure 23.23 Enzymatic cleavage of *N*-glycans of glycoproteins with endoglycosidases (Endo-H, Endo-F) or glycopeptidases (PNGase F, PNGase A).

Enzymatic Release by Endoglycosidases *N*-Glycan release via endoglycosidase consists of hydrolysis of the chitobiose disaccharide unit (Figure 23.23). Here, essentially four endoglycosidases are available; their specificities are summarized in Table 23.7. Endo-H and Endo-F_1 cleave *N*-glycans of the *high-mannose* and hybrid type, but not the *complex* type structures. Thus, the *high-mannose* type and *hybrid* type structures of glycoproteins can selectively be released (best as subsequent to tryptic digestion), while the *complex* type *N*-glycans remain bound to the peptide backbone. These may then be partially be cleaved with Endo-F_2 or Endo-F_3 or completely cleaved with PNGase F or PNGase A.

Chemical Release by Hydrazinolysis In hydrazinolysis that requires relatively harsh conditions (incubation for several hours in anhydrous hydrazine at 80–100 °C), the *N*-glycosidic amide bond is hydrolyzed to asparagine, whereby several conversions take place. The *N*-glycans, however, will ultimately be isolated in intact, reducing form – as in the enzyme reaction with PNGase F. The chemical process is, however, only rarely used because of the difficult access to anhydrous hydrazine and the associated toxicity and is therefore not described in detail.

Table 23.7 Endoglycosidases and their respective cleavage specificity.

	High-mannose type	Hybrid type	Complex type
Endo-H	+	+	–
Endo-F_1	+	+	–
Endo-F_2	(+)	–	Biantennary (+/- *core*-Fuc)
Endo-F_3	–	–	Biantennary + *core*-Fuc (biantennary without *core*-Fuc) (triantennary without *core*-Fuc) Trimannosyl + *core*-Fuc

23.3.3.1 Analysis of the *N*-Glycan Pool

As mentioned above, the consistency of biopharmaceutical products (and thus the glycosylation) must be checked from batch to batch. For this purpose several analytical techniques are available, the selection of which depends on the complexity of the glycoprotein, the influence of glycosylation on the biological activity, the pharmaceutical safety of the product, and the general strategy of manufacturing control. Even if the importance of glycosylation for therapeutic efficacy of the product is not fully ensured, batch-to-batch consistency of the product may be checked and evaluated via assessment of glycosylation.

Batch consistency of glycosylation is generally covered by isolation and chromatographic analysis of the glycan pool. This so-called *N*-glycan mapping gives insight into the various populations of the glycans and the degree of sialylation. It may be performed by HPAEC-PAD using the native oligosaccharides or by HPLC – after derivatization of the glycans at the reducing end (introduction of a UV or fluorescent label).

The hypothetical charge number *Z* thus allows us to characterize the glycosylation of a glycoprotein in a simple manner and therefore represents an important parameter of safety pharmacology.

Mapping of Native *N*-Glycans Anion-exchange chromatography (HPAEC-PAD/generally: matrix with tertiary amine) separates the glycans primarily according to charge into peak groups with neutral or asialo-(AS), monosialo (MS), disialo (DiS), trisialo (TriS), and tetrasialo (TetraS) *N*-glycans. This mapping profile may be described quite well in the form of a single number – the hypothetical charge number Z (*Z-number*). For this the chromatographic software records the percentages (*A*) of the different peak groups (of uniform charge). The percentage share of each peak group is then multiplied by its corresponding charge and the obtained products (charge shares) are finally summarized:

$$Z = (A_{(AS)}0) + (A_{(MS)}1) + (A_{(DiS)}2) + (A_{(TriS)}3) + (A_{(TetraS)}4) + (A_{(PentaS)}5) \qquad (23.1)$$

Thus, a glycoprotein which has mainly tetraantennary tetrasialo structures should provide a charge number of about 400; a glycoprotein having predominantly triantennary trisialo

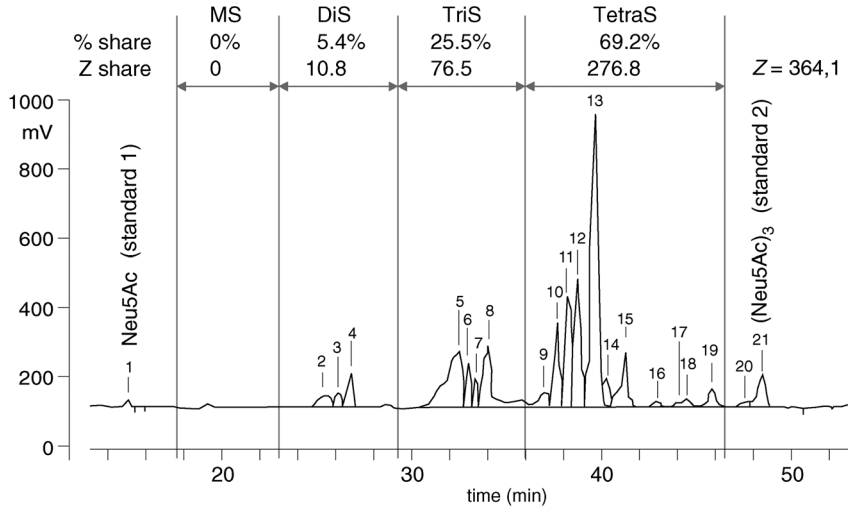


Figure 23.24 HPAEC-PAD mapping profile of the native *N*-glycan pool of rhEPO (CHO) using standard gradient "S" (Table 23.9) and determination of the hypothetical charge number *Z* (*Z-number*); the structural assignment of the *N*-glycans via a database may be found in Table 23.10. As can be seen from the mapping profile, a total of seven peaks is detected in the peak group of the normal tetra-sialylated *N*-glycans (between 36 and 42 min). The main peak is the tetrasialo-(α2,3)-tetra-antennary *N*-glycan with proximal fucose (4N(2,3)-A4, peak 13). The second largest peak (no. 12) has the same structure, but with an additional antennary LacNAc *repeat* (4N(2,3)-A4+1R). Peak 11 corresponds to the isomer of peak 12 in which the LacNAc-*repeat* is located on the uppermost (instead of the second uppermost) antenna. Peak 10 is the glycan with two LacNAc *repeats* (4N(2,3)-A4+2 R), corresponding to peak 13, where one *repeat* each is located on the uppermost and the second uppermost antenna. Under the strongly alkaline mapping conditions of HPAEC-PAD, a certain percentage of the GlcNAc at the reducing end of the glycans (the proximal GlcNAc) may epimerize to ManNAc (*N*-acetylmannosamine). The ManNAc epimer to peak 13 is detected as peak 15, while peak 14 is the ManNAc epimer of peak 12. The small peaks (spikes) 16–19 are derived from glycan structures 10–13 by replacing one Neu5Ac residue for one Neu5Gc residue, each, which causes an increase in the retention time of each by approximately 6 min. (rhEPO (CHO) – in contrast to rhEPO (BHK) – contributes 1–3% Neu5Gc.)
In the trisialo range (peak group between 30 and 36 min), the trisialo-(α2,3)-tetra-*N*-glycan corresponding to peak 13 is detected as peak 6. Peak 8 reflects the two trisialo-(α2,3)-triantennary isomers, which, under the measurement conditions, cannot be separated into two peaks. In the peak group of the disialo range (24–28 min), the disialo-(α2,3)-tetra-*N*-glycan corresponding to peak 13 is detected as peak 3. Peak 4 corresponds to the disialo-(α2,3)-biantennary *N*-glycan.

Table 23.8 Hypothetical charge number *Z* of some glycoproteins. Source: according to Hermentin, P, *et al.* (1996) *Glycobiology*, **6**, 217–230.

Glycoprotein	Manufacturer	Origin of glycan pool	*Z*
rhu EPO (CHO)	Böhringer, Mannheim	PNGase F	361
rhu EPO (CHO)	Amgen	PNGase F	367
rhu EPO (CHO)	Organon Teknika	PNGase F/SDS	286
rhu EPO (BHK)	Merckle	PNGase F	323
Bovine fetuin	Sigma	Hydrazinolysis	290
Bovine pancreas ribonuclease B		OGS[a]	15
Hen ovomucoid		OGS	15
Pig thyroglobulin		OGS	82
Human α-1 acid glycoprotein	BW AG[b]	Hydrazinolysis	289
Human serotransferrin		OGS	207
Human antithrombin III	BW AG	Hydrazinolysis	180
Human fibrinogen		OGS	184
α-1-T-glycoprotein	BW AG	Hydrazinolysis	187
α-1-Antitrypsin	BW AG	Hydrazinolysis	190
α-1-Antichymotrypsin	BW AG	Hydrazinolysis	236
β-2-Glycoprotein I	BW AG	Hydrazinolysis	185
TBG-glycoprotein	BW AG	Hydrazinolysis	208
α-1-B-glycoprotein	BW AG	Hydrazinolysis	194
α-2-HS-glycoprotein	BW AG	Hydrazinolysis	158
8S-α-3-glycoprotein	BW AG	Hydrazinolysis	145
Haptoglobin	BW AG	Hydrazinolysis	197

structures would have a charge number of about 300; for a glycoprotein with essentially biantennary disialo glycans a charge number of ca. 200 may be calculated; and for a glycoprotein that exhibits only *high-mannose* type or *truncated* *N*-glycan structures the charge number should practically be zero. Thus, the simple numerical magnitude of the hypothetical charge number *Z* indicates the type and weighting of the *N*-glycans of a glycoprotein.

Thus, the HPAEC-PAD mapping profile of the rhEPO *N*-glycan pool (CHO, Figure 23.24) allows to calculate a charge number of:

$$Z = (0 \times 0) + (0 \times 1) + (5.4 \times 2) + (25.5 \times 3) + (69.2 \times 4) = 364.1$$

Further examples are listed in Table 23.8.

In addition, the mapping profile allows a structural assignment of individual peaks based on a simple comparison of the measured retention times with those of a mapping database or of standard glycans of known structure.

The separation of the native N-glycan pool of EPO (CHO) by means of HPAEC-PAD (standard gradient "S," Table 23.9) is shown in Figure 23.24, while the structural assignment via a database is summarized in Table 23.10.

An alternative is to separate the reduced *N*-glycan pool of EPO (CHO) via a graphitized carbon column, GCC (Figure 23.29 below) and assign the structure via ESI-MS (Figure 23.30 below).

Desialylation of the *N*-glycan pool by neuraminidase and subsequent analysis by HPAEC-PAD yields a separation of the asialo-*N*-glycans (Figure 23.25). From this it is obvious that HPAEC-PAD separates the neutral glycans primarily according to their size or antennarity. Integration of the individual peaks approximately reflects the percentage share of each of the neutral structures of the total *N*-glycan pool.

For the structural assignment of *N*-glycans via HPAEC-PAD, the following rules are helpful:

- The loss of one, two, three, or four neuraminic acid residues shortens the retention time in HPAEC-PAD (using the standard gradient "S" for sialylated glycans described in the literature, see Table 23.9) by 6.5, 13, 19.5, and 26 min. (This shows the strong influence of negatively charged Neu5Ac residues on the interaction of the glycan with the anion-exchange matrix.)
- The introduction of an antennary LacNAc repeat shortens the retention time of sialylated *N*-glycans, while retention time of the corresponding asialo-*N*-glycans will be increased.
- Glycans with α2,3-linked Neu5Ac residues always elute later than their corresponding sialyl isomers with α2,6-linked Neu5Ac.
- Glycans fucosylated at the proximal GlcNAc always elute earlier than their non-fucosylated analogues.
- Replacement in an *N*-glycan of Neu5Ac by Neu5Gc (*N*-glycolylneuraminic acid) causes a significant increase of retention time.
- Glycan sulfation results in a dramatic increase of retention time, compared to the non-sulfated analogue, because the strong polar sulfate group causes a particularly intense interaction between glycan and anion-exchange matrix.

Table 23.9 Standard gradient "S" for sialylated *N*-glycans. Source: according to Hermentin, P, *et al.* (1992) *Anal. Biochem.*, **203**, 281–289.

Gradient	Elution time (min)	Eluent 1 (0.1 M NaOH)	Eluent 2 (0.6 M NaAc in 0.1 M NaOH)
S	0	100	0
	2	100	0
	50	65	35
	57	30	70
	60	0	100
	63	0	100
	64	100	0
	70	100	0

Table 23.10 rhuEPO (CHO) *N*-glycan structural assignment via comparison of the HPAEC-PAD mapping profile shown in Figure 23.24 with a database.

Peak number	Retention time (min)	*N*-Glycan component (each with proximal fucose)	Alternative designation[a]	Designation in the database
1	14.97	Neu_5Ac (standard 1)		
2	25.28			
3	26.00	Disialo tetraantennary	$TetraNA_2$	C4-246301.20
4	26.65	Disialo biantennary	$BiNA_2$	C2-224301.20
5	32.39	Trisialo tetraantennary + 1R	$TeraLac_1NA_3$	C4-357301.30.1R
6	32.90	Trisialo tetraantennary	$TetraNA_3$	C4-346301.30
7	33.30			
8	33.95	Trisialo triantennary	$TriNA_3$	C3-335301.30
9	36.96			
10	37.60	Tetrasialo tetraantennary + 2 R	$TetraLac_2NA_4$ (e)[b]	C4-468301.40.2R
11	38.25	Tetrasialo tetraantennary + 1 R (isomer I)	$TetraLac_1NA_4$ (II) (d)[b]	C4-457301.40.1R-A6M6
12	38.68	Tetrasialo tetraantennary + 1 R (isomer II)	$TetraLac_1NA_4$ (I) (c)[b]	C4-457301.40.1R-A2M6
13	39.57	Tetrasialo tetraantennary	$TetraNA_4$ (b)[b]	C4-446301.40
14	40.23	ManNAc epimer of peak 12		
15	41.27	ManNAc epimer of peak 13		
16	42.96	Neu_5Gc-isomer of peak 10		
17	44.07	Neu_5Gc-isomer of peak 11		
18	44.51	Neu_5Gc-isomer of peak 12		
19	45.83	Neu_5Gc-isomer of peak 13		
20	47.61			
21	48.47	$(Neu_5Ac)_3$ (standard 2)		

C4-446301.40
- 1 *Repeat*
- 0 NeuAcα2,6
- 4 NeuAcα2,3
- 1 *proximal* Fuc
- 0 *bisecting* GlcNAC
- 3 Man
- 6 GlcNAc
- 4Gal
- 4 Neu5Ac
- 4 Antennen

Complex Type

a) Designation according to Kawasaki, N. *et al.* (2000) *Anal. Biochem.*, **285**, 82–91.
b) Corresponding reduced structures of Figure 23.29.

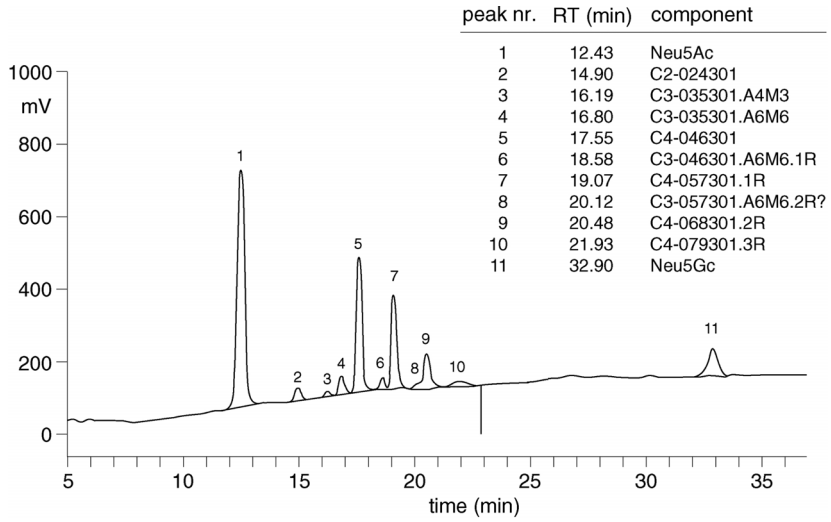


Figure 23.25 HPAEC-PAD mapping profile of the asialo-*N*-glycan pool of rhEPO (CHO). By desialylation of the glycan pool, the differently sialylated *N*-glycans of peaks 13, 6, and 3 from Figure 23.24 are transferred into one and the same tetraantennary asialo-*N*-glycan (main peak, number 5). Peak 7 corresponds to the tetraantennary asialo-*N*-glycan enlarged by one LacNAc-*repeat*, resulting from the isomer peaks 11 and 12 of the sialylated *N*-glycans from Figure 23.24. Peaks 9 and 10 correspond to the tetraantennary asialo-*N*-glycans with two or three LacNAc-*repeats*. The introduction of one, two, or three LacNAc-*repeats* thus causes for an asialo-*N*-glycan a significant increase in the retention time, which is due to an increased interaction of the glycan with the column matrix. The two triantennary glycan isomers not separated in the sialylated form (peak 8 of Figure 23.24) are, after desialylation, separated into two distinct peaks that are about 0.6 min apart (peaks 3 and 4). The biantennary asialo-*N*-glycan finally elutes with the shortest retention time (peak 2).

Mapping of Derivatized *N*-Glycans

Derivatization of *N*-Glycans at the Reducing End Derivatization at the reducing end is intended to make the glycans, which absorb only weakly in the range of 190–210 nm, more sensitive for detection and (if necessary) to provide neutral glycans with a charge.

Thereby, one takes advantage of the fact that the reducing sugar unit in its open-chain aldehyde form allows a Schiff base reaction with a primary amine. Thus, *N*-glycans may be labeled in a specific manner with different UV or fluorescence-active groups (Figure 23.26). Figure 23.27 exhibits the most common fluorescence-active groups.

Here, especially APTS with three sulfate groups and a highly fluorescent pyrene ring has proven its usefulness in capillary electrophoresis with laser-induced fluorescence (CE-LIF), which enables quick and high-resolution analysis. CE has prevailed especially in the analysis of *N*-glycans from antibodies. As 2-AA derivatives carry a lower negative charge than APTS derivatives, their analysis is comparably longer, and the separation performance of the CE is inferior. However, in the analysis of monoclonal antibodies, the 2-AA method has the advantage that the sialylated and *high-mannose* structures are completely captured, and the *N*-glycans fucosylated at the proximal GlcNAc are completely separated from their non-fucosylated analogues (the non-fucosylated species exhibit shorter migration times).

Two-Dimensional Mapping of Pyridylaminated *N*-Glycans A commonly used derivatization is the reaction with 2-aminopyridine (2AP, Figure 23.26). This process, developed and

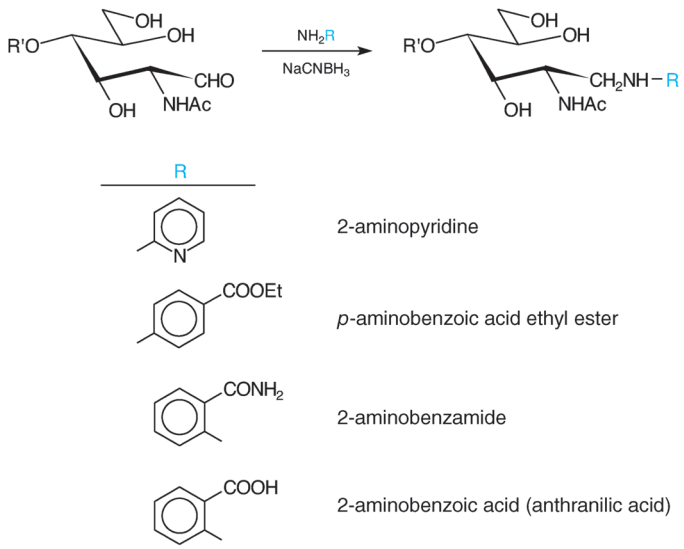


Figure 23.26 Labeling of an *N*-glycan at the reducing end with 2-aminopyridine, ethyl *p*-aminobenzoate, 2-aminobenzamide, or 2 aminobenzoic acid (anthranilic acid); R = oligosaccharide residue.

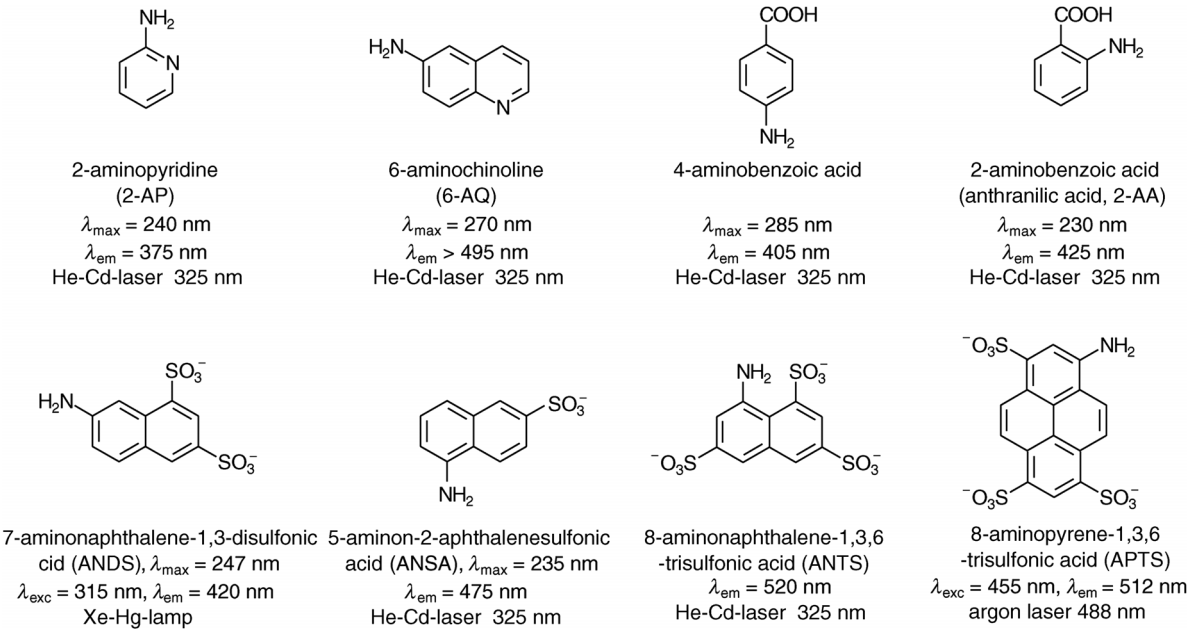


Figure 23.27 *N*-Glycan derivatives for CE. Under acidic electrophoresis conditions, 2-aminopyridine and 6-aminoquinoline are positively charged. *p*-Aminobenzoic acid, 2-AA, ANDS, ANTS, and APTS are used as negatively charged markers.

favored by Japanese researchers, has over the years been modified and optimized several times. The reduction of the Schiff base formed between glycan and 2-AP today is best performed with a dimethylamine-borane complex (instead of $NaCNBH_3$) because this reagent, on the one hand, ensures the preservation of the Neu5Ac residues and, on the other hand, may relatively easily be removed from the reaction mixture.

The pyridylaminated (PA) oligosaccharides are analyzed on two different HPLC columns. In this 2D-mapping procedure the desialylated PA-oligosaccharides are separated on an ODS (octadecylsilyl) silica column and a TSK Amide-80® column, and their respective elution volumes are determined as glucose units (GU) in relation to an isomaltose oligosaccharide standard. The structural assignment of the *N*-glycans is finally made with the use of a database of more than 400 pyridylaminated oligosaccharides (status 2002) by simple comparison of the precisely measured elution volumes. In combination with enzymatic sequencing, an unknown *N*-glycan can often be converted into a known structure contained in the database and thus be structurally assigned (Figure 23.28).

Depending on the type of derivatization and the nature of the glycan pool, IE (ion-exchange), RP (reversed phase), NP (normal phase), or GCC (graphitized carbon column) HPLC may here

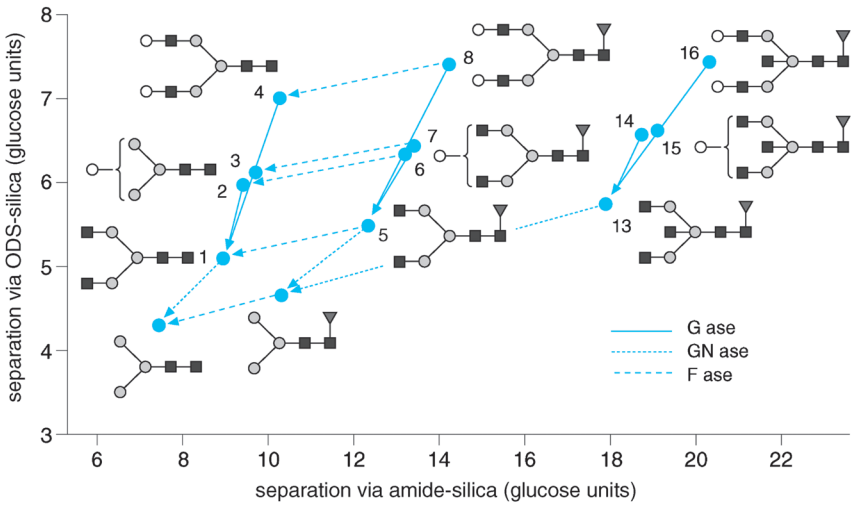


Figure 23.28 Two-dimensional mapping of PA-oligosaccharides. The biantennary *N*-glycan number 8 is converted into glycan number 4 (upon cleavage of the proximal fucose residue by α-L-fucosidase; F ase), which can be further degraded (via the two monogalactosyl isomers 2 and 3) to glycan (1) (removal of the two terminal galactose residues by β-galactosidase; G ase). Glycan (1) eventually will be converted into the trimannosyl-pentasaccharide *core* (by means of the enzyme *N*-acetylglucosaminidase; GN ase).
Analogously, glycan number 16 may, via the two monogalactosyl isomers 14 and 15, can be converted into glycan 13 (through β-galactosidase). This is converted (by means of the enzyme GN ase) into the fucosylated trimannosyl pentasaccharide *core*, which eventually can be converted into $Man_3GlcNAc_2$ (via fucosidase). The changes in the elution volumes observed in the enzymatic degradation reactions (relative to an isomaltose oligosaccharide standard) may be read via a 2D map (abscissa: ODS-silica (ODS = octadecylsilyl); ordinate: amide-silica). Source: according to: Lee, Y.C. and Rice, K.G. (1995) in *Glycobiology: A Practical Approach* (eds M. Fukuda and A. Kobata), ch. 3C, IRL Press.

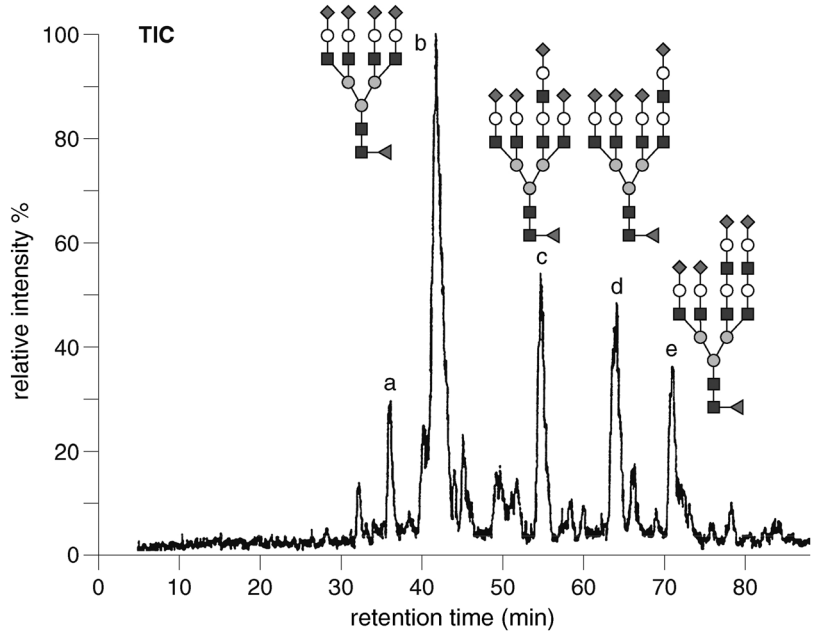


Figure 23.29 Relative intensities of the reduced *N*-glycans of EPO (CHO) via ESI-MS (negative ion mode) and presumed structural assignment after separation of the reduced *N*-glycan pool on a graphite column. Source: according to Kawasaki, N. *et al.* (2000) *Anal. Biochem.*, **285**, 82–91. With permission, Copyright © 2000 Academic Press.

likewise be considered. Structural assignment of the glycans separated by HPLC can be ensured by comparison with entrained glycan standards or by coupling with MS. Thus, RP-HPLC of the glycans in combination with ESI-MS(MS) enables an online analysis that combines the qualitative and quantitative mapping profile with information on the glycan structure (molar mass, fragment pattern) in a single run.

Mass Spectrometric Analysis of the *N*-Glycan Pool In the characterization of glycans, mass spectrometry (ESI-MS or MALDI-MS) is nowadays practically indispensable. For this, glycan separation via HPAEC-PAD requires neutralization of the alkaline eluent and desalting via a membrane module, whereas IE or NP-HPLC need desalting via a suitable solid phase extraction or dialysis module. Only in the case of RP-HPLC and GCC-HPLC are the glycans eluted virtually salt-free, which enables direct LC/MS coupling.

Figure 23.29 exhibits the separation of the reduced *N*-glycan pool of rhEPO (CHO) via GCC-HPLC with online analysis of the detected masses via an ESI mass spectrometer (eluent: 5 mM ammonium acetate in 50% acetonitrile). The mass spectra of the main peaks a–e are shown in Figure 23.30. The main peak b is the tetrasialo tetraantennary *N*-glycan ($TetraNA_4^{3-}$), peaks c and d are the two tetrasialo tetraantennary *N*-glycan isomers with one LacNAc *repeat* ($TetraLac_1NA_4^{3-}$), and peak e is the tetrasialo tetraantennary *N*-glycan with two LacNAc *repeats* ($TetraLac_2NA_4^{3-}$). The peak pattern is similar to the one of HPAEC-PAD (Figure 23.24), but there the order is exactly reversed (peak 13 = b, peak 12 = c, peak 11 = d, and peak 10 = e).

Mass Spectrometry, Chapter 15

Structural assignment of the two isomers c and d (i.e., an allocation of the single LacNAc *repeat* to a particular antenna), however, is not possible on the basis of an HPLC/MS study. In addition, methylation analysis and exoglycosidase digestion (Section 23.3.5) are also not useful here. This question can only be clarified via relatively complex ^{1}H NMR studies (Section 23.3.5).

Analysis of Individual Glycosylation Sites In Section 23.3.2 the *site-specific glycosylation* was analyzed on the basis of the glycopeptides separated via RP-HPLC and the glycan structures per glycosylation site was assigned by LC/MS (cf. Figure 23.21 and Table 23.6). If one now wants to deepen these studies on the basis of the free *N*-glycans, one has initially to fractionate the individual RP-HPLC peaks. The thus isolated glycopeptides should then be digested with PNGase, and the released glycan pools should each be desalted (gel filtration, e.g., Sephadex, or solid phase extraction, e.g., Hypercarb^{TM}). The isolated *N*-glycan pools are each separated – either in native form via HPAEC or, after derivatization with a UV or fluorescence-active group, via HPLC (IE, RP, NP, GCC) – and the individual *N*-glycan peaks are measured via mass spectrometry either online (LC/MS) or after fractionation offline (ESI-MS, MALDI-MS). As mentioned above, the configuration of the *N*-glycans may here be assigned mostly via their elution

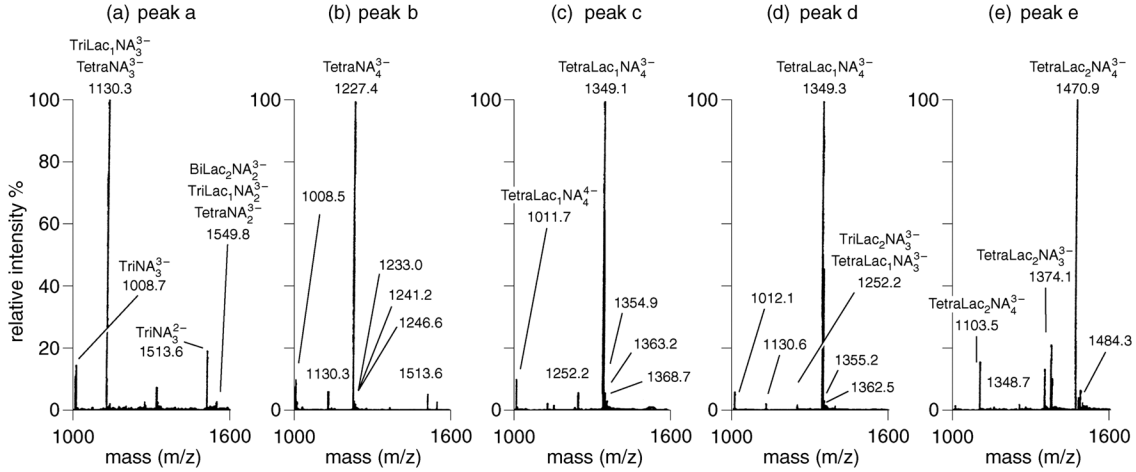


Figure 23.30 Mass spectra of peaks a–e of Figure 23.29 (ESI-MS, negative ion mode. The tetrasialylated *N*-glycans of the mass spectrum are largely triple-negatively charged $[M\text{-}3H]^{3-}$, the doubly or quadruply charged ions are only weakly detected. Source: according to Kawasaki, N. *et al*. (2000) *Anal. Biochem*., **285**, 82–91. With permission, Copyright © 2000 Academic Press.

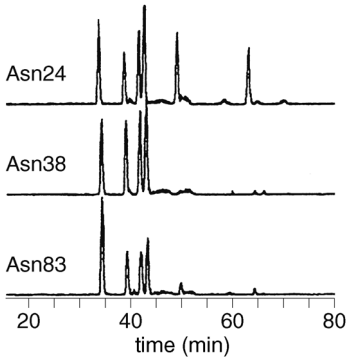


Figure 23.31 GCC/ESI-MS of the reduced *N*-glycan pool of Asn24, Asn38, and Asn83 of rEPO (CHO) recorded in the *SIM scan* mode at $m/z = 1130$, corresponding to the trisialo isomers $TriLac_1NA_3$ (trisialo-triantennary with 1 LacNAc-*repeat*) or $TetraNA_3$ (trisialo-tetraantennary). Source: according to Kawasaki, N. *et al*. (2000) *Anal. Biochem*., **285**, 82–91. With permission, Copyright © 2000 Academic Press.

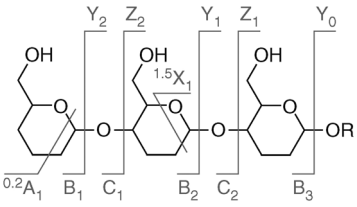


Figure 23.32 Nomenclature of fragment ions. Source: according to Henning, S. *et al.* (2009) in *Mass Spectrometry of Proteins and Peptides* (eds M. Lipton and L. Pasa-Tolic), vol. 492, Humana Press, ch 10, pp. 181–200. With permission, Copyright © 2009, Humana Press, a part of Springer Science+Business Media, LLC.

behavior in HPAEC and HPLC in comparison with reference standards measured in parallel or by comparisons with an appropriate database (Section 23.3.4). The differences in glycosylation per *site* can easily be assigned by comparison of the chromatograms (Figure 23.31).

Mass Spectrometric Sequencing Free or peptide-bound sugar chains may be fragmented into various cleavage products and sequenced, like peptide chains (cf. Section 16.1.5), using tandem MS (MS/MS) experiments. Glycosidic bonds are generally weaker than peptide bonds and therefore fragment more easily. If during fragmentation the charge remains with the terminally cleaved sugar fragments, these are indexed B_i. The terminal monosaccharide fragment ion is termed B_1, the terminal disaccharide fragment ion B_2, the terminal trisaccharide fragment ion B_3, and so on. If the generated fragment ions consist of the reducing end or the peptide residue, they are denoted Y_i. The reducing or Asn-linked monosaccharide group is designated Y_1, the corresponding disaccharide residue Y_2, the corresponding trisaccharide unit Y_3, and so on (Figure 23.32) – in analogy to the fragmentation of a peptide chain in b/y fragments (cf. Section 15.5, Figure 15.33). In biantennary glycans the antennary fragments are indexed with the suffix α or β. A ring-cleavage is referred to as A_i (non-reducing terminus) and X_i (reducing terminus), wherein the two cleaved bonds are each identified by a preceding superscript index that characterizes the cleaved bonds (Figure 23.32).

Thus, the fragment ion spectrum of the biantennary *N*-glycan from human transferrin recorded by means of nano-ESI Q-TOF shows almost the entire series of the Y and B ions and other characteristic fragments that result from a double-fragmentation (Figure 23.33). Moreover, a few, but characteristic, b and y ions of the peptide backbone show that the studied glycopeptide fragments are fragments of amino acid sequence 421–446, which is glycosylated at N432.

The molecular ion (m/z 1264.04) is detected as a quadruply charged species. The monosaccharide ($B_{1\alpha}$), disaccharide ($B_{2\alpha}$), trisaccharide ($B_{3\alpha}$), and tetrasaccharide ($B_{4\alpha}$) fragments cleaved from the nonreducing end correspond to the antennary glycan sequence, which may also be deduced from the corresponding glycopeptide fragments $Y_{6\alpha}$, $Y_{5\alpha}$, $Y_{4\alpha}$, and $Y_{3\alpha}$ detected in the spectrum as triply charged species. Fragment ion B_6 belongs to the double-charged, biantennary decasaccharide fragment, fragment ion Y_1 to the corresponding peptide fragment containing the Asn-linked GlcNAc residue. Fragments $Y_{6\beta}/B_{3\alpha}^{3+}$, $Y_{5\beta}/B_{3\alpha}^{3+}$, $Y_{4\beta}/B_{3\alpha}^{2+}$, and $Y_{3\beta}/B_{3\alpha}^{2+}$ correspond to the glycopeptide fragments in which on one antenna (β) a mono- or di- or trisaccharide fragment was cleaved and on the second antenna (α) the terminal trisaccharides $B_{3\alpha}$ were missing (double fragmentation). The terminal Neu5Ac fragment $B_{1\alpha}$ is, after the loss of water, also detected as $B_{1\alpha} - H_2O$. The very complex mass analysis is performed by appropriate computer programs.

However, mass spectrometry does not reveal the anomeric configuration of the individual sugar units in the glycans, nor their respective linkage direction. A complete analysis of glycosylation therefore also requires methylation analysis, exoglycosidase digestions, and (as *ultima ratio*) ^1H NMR studies of the isolated glycans (Section 23.3.5).

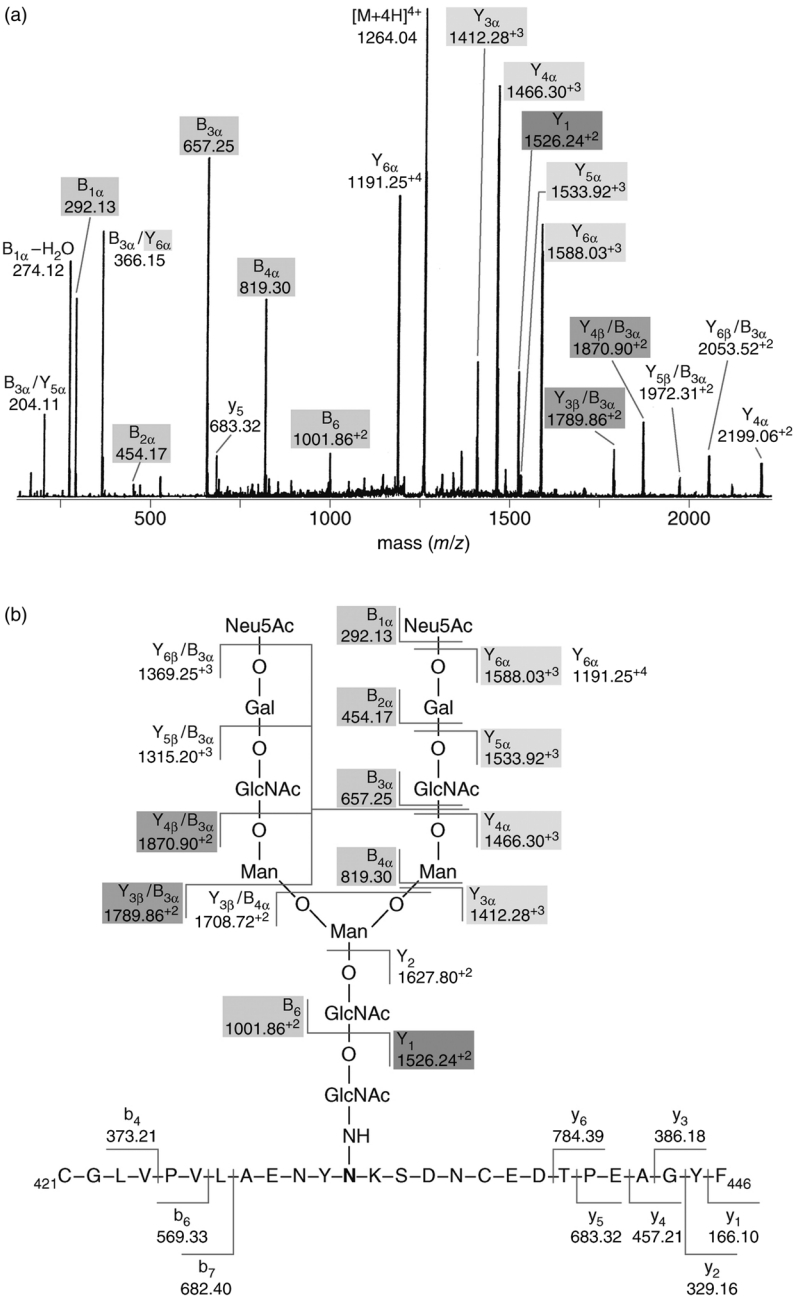


Figure 23.33 (a) Fragment-ion spectrum of the quadruply charged precursor ion of $m/z = 1264.04$ of tryptically/chymotryptically digested human transferrin (nanoESI Q-TOF, CID). (b) Assignment of the fragment-ions shown in (a). Source: according to Henning, S. *et al.* (2009) in *Mass Spectrometry of Proteins and Peptides* (eds M. Lipton and L. Pasa-Tolic), vol. 492, Humana Press, ch 10, pp. 181–200. With permission, Copyright © 2009, Humana Press, a part of Springer Science+Business Media, LLC.

23.3.4 Analysis of Individual *N*-Glycans

In the following we turn our attention to the structure elucidation of isolated individual glycans, which is a relatively difficult task. In particular, the important techniques of mass spectrometry and ^{1}H NMR spectroscopy make high demands on the equipment and on scientific expertise and therefore only some important basics are discussed in the following subsections.

Isolation of Individual *N*-Glycans Purification of individual *N*-glycans from the *N*-glycan pool is quite labor-intensive. Various working groups have developed different separation schemes (with or without derivatization of the glycans at the reducing end) that have been optimized for their specific laboratory conditions.

The easiest way to isolate the non-derivatized (intact) glycans is to follow the separation scheme shown in Figure 23.14 (left-hand column). In the first step, the *N*-glycans are separated

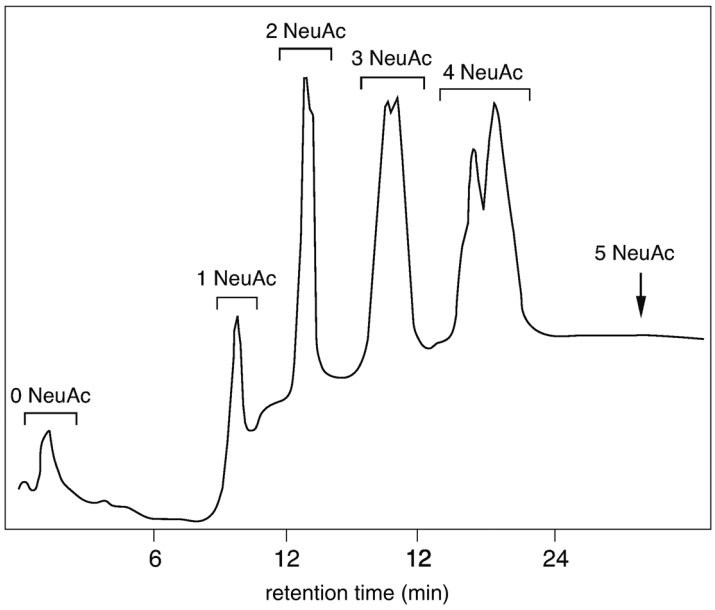


Figure 23.34 Mono-Q separation of the *N*-glycan pool of recombinant erythropoietin from BHK cells. As can be seen from the figure, rEPO (BHK) exhibits *N*-glycans carrying 1–4 Neu5Ac residues, wherein the weighting of the glycans increases with increasing charge. Source: according to Nimtz, M. *et al*. (1993) *Eur. J. Biochem*., **213**, 39–56. With permission, Copyright © 2005, John Wiley and Sons.

via an anion-exchange matrix (Mono-Q) into glycans of identical charge (same sialylation) (Figure 23.34 for rhEPO (BHK)). This allows us to determine the sialylation status of the glycoprotein, that is, the percentage composition of the *N*-glycan pool of neutral or asialo (AS), monosialo (MS), disialo (DiS), trisialo (TriS), and tetrasialo (TetraS) structures. The glycans are detected at short wavelength UV (190–210 nm) – here the C═O bonds of the *N*-acetyl groups of GlcNAc and Neu5Ac and the carboxy groups of Neu5Ac residues absorb.

The individual Mono-Q fractions (MQ-1, MQ-2, MQ-3, and MQ-4) may in a further chromatographic step via an amino phase (*LiChrosorb*-NH_2) be separated into individual glycans and again be detected at 190–210 nm (Figure 23.35 for the separation of MQ-4 of rhEPO (BHK)). This results in a set of *N*-glycan peaks that are collected via a fraction collector and desalted through Sephadex G-15.

Each of these peaks (*N*-glycans) is then checked with respect to its purity via HPAEC-PAD, which has a greater selectivity than Mono-Q and *LiChrosorb*-NH_2, and is finally subjected to a detailed structural analysis that (depending on the strategy) may consist of composition analysis, methylation analysis, exoglycosidase digestion, mass spectrometry, or high-resolution ^{1}H NMR spectroscopy (see below).

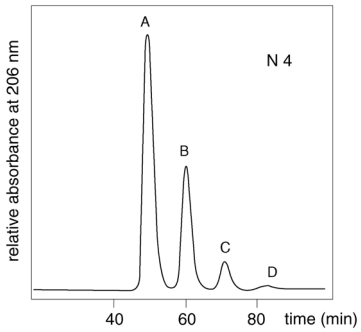


Figure 23.35 Separation of Mono-Q fraction MQ-4 of rEPO (BHK) via *LiChrosorb*-NH_2. Amino phase separation (N4) of tetrasialo fraction MQ-4 of rEPO (BHK) results in a main peak A and three secondary peaks B–D. Source: according to Nimtz, M. *et al*. (1993) *Eur. J. Biochem*., **213**, 39–56. With permission, Copyright © 2005, John Wiley and Sons.

Composition Analysis For analysis of the monosaccharide components of isolated glycans (composition analysis), in principle the same analytical methods apply as have already been described in Section 23.3.1 for the intact glycoprotein. The analysis of isolated glycans also initially requires the total hydrolysis of all glycosidic linkages (e.g., in the presence of 2 N TFA at 100 °C within 4 h), whereby the *N*-acetylated sugars (GlcNAc and GalNAc) will be deacetylated and converted into the corresponding amino sugars ($GlcNH_2$ and $GalNH_2$) and detected as such.

Gas chromatographic analysis on the basis of the monosaccharide alditol acetates formerly used in combination with a mass spectrometer (GC-MS) is nowadays only rarely carried out because this information is more easily accessible via monosaccharide analysis via HPAEC-PAD (Section 23.3.1, Figure 23.19).

Methylation Analysis Methylation analysis is the oldest method still routinely used to determine the linkage direction between the individual monosaccharide components of a glycan. However, it provide no information on the sequential linkage and anomery, that is, the respective monosaccharide configuration at C1.

Methylation analysis of an isolated *N*-glycan consists of the following steps (Figure 23.36):

1. Reduction of the *N*-glycan to the alditol.
2. Permethylation of the free OH groups.

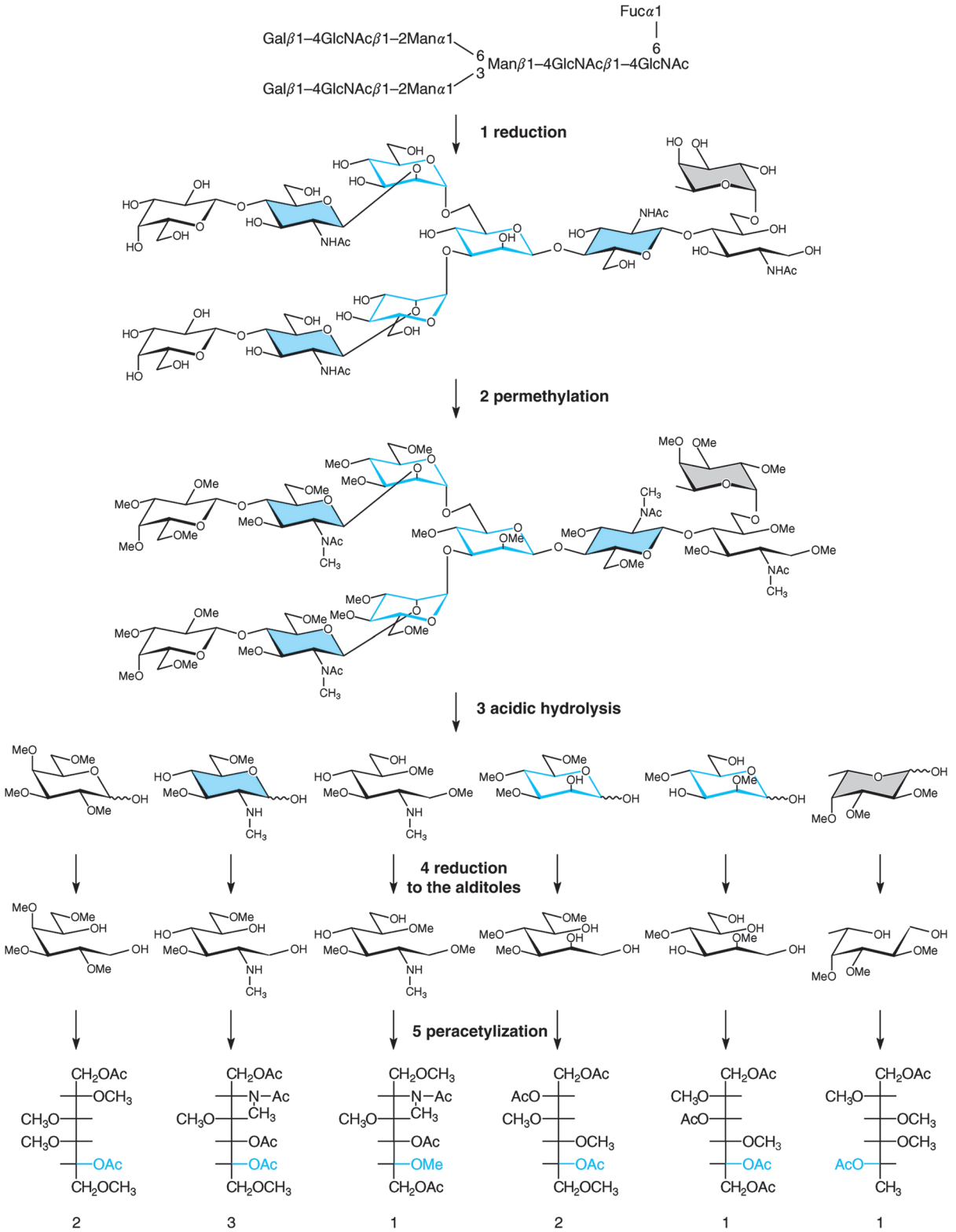


Figure 23.36 Methylation analysis of the biantennary asialo-*N*-glycan with proximal fucose. Analysis after total hydrolysis of the reduced, permethylated structure yields five pairs of partially methylated monosaccharide anomers and (after reduction to the corresponding alditol) the respective six partially methylated alditols (two equivalents of 2,3,4,6-tetra-*O*-methyl-D-galactitol, three equivalents of 2-deoxy-2-*N*-methylacetamido-3,6-di-*O*-methyl-D-glucitol, one equivalent of 2-deoxy-2-*N*-methylacetamido-1,3,5-tri-*O*-methyl-D-glucitol, two equivalents of 3,4,6-tri-*O*-methyl-D-mannitol, one equivalent of 2,4-di-*O*-methyl-D-mannitol, and one equivalent of 2,3,4-tri-*O*-methyl-L-fucitol). Peracetylation of the partially methylated alditols yields the corresponding methylated alditol acetates each in a molar ratio of 2 : 3 : 1 : 2 : 1 : 1.

3. Acid hydrolysis of all glycosidic bonds: thereby, the free monosaccharide building blocks arise as derivatives in which the OH groups that in the *N*-glycan have previously been free are now marked by a methyl group.
4. Reduction of the methylated monosaccharide to alditols: this converts the partially methylated monosaccharides that in each occur as anomeric pairs (with α/β-OH) into the corresponding alditols, whereby the anomeric problem is eliminated.
5. Peracetylation (of the remaining free OH groups): the acetyl groups present in the methylated alditol acetates thus mark in each monosaccharide building block the OH group at C1 (with the exception of the proximal GlcNAc) as well as those OH groups that previously (in the intact glycan) have been linked in a glycosidic bond.
6. Measurement of the permethylated alditol acetates by GC-MS: this enables the identification and quantification of the individually permethylated alditol acetates by comparison with commercially available synthetic standards.

The methylation analysis is illustrated, by way of example, using a biantennary asialo-*N*-glycan (Figure 23.36). It provides important information about the connection points of each of the monosaccharide building blocks. When assembling the units (setting the order of linkage direction and anomery) one has to rely on additional analysis methods, such as enzymatic sequencing and high-resolution ^{1}H NMR spectroscopy (see below). In standard mammalian cell expression systems with well-known glycosylation capacity (e.g., BHK or CHO cells), the expected glycan structures can, however, already be predicted with good probability from the data of methylation analysis.

Exoglycosidase Digestion Exoglycosidases are enzymes that split from the nonreducing end of an oligosaccharide exactly the one monosaccharide unit that corresponds in its anomeric configuration and its linkage direction with the specificity of the respective enzyme. For example, α(1,3)-galactosidase cleaves from a given *N*-glycan only terminal α(1,3)-linked galactose residues, while terminal β(1,4)-linked galactose residues remain preserved.

This allows to analyze a glycan with a series of well-defined exoglycosidases, used either sequentially or in parallel (in array), by enzymatic sequencing and to assign its configuration (with respect to anomery and linkage direction) (Table 23.11). The degradation may be monitored by HPAEC-PAD (non-derivatized *N*-glycans, Figure 23.37) or by RP-HPLC (Figure 23.28) or mass spectrometry (Figure 23.38) (derivatized *N*-glycans). The degradation of fluorescent-labeled glycans may also be monitored in an agarose gel (fluorophore-assisted carbohydrate electrophoresis, FACE) – here the labeling of the *N*-glycans with 8-amino-naphthalene-1,3,6-trisulphonic acid (ANTS, Figure 23.27) has proven successful.

Nowadays, sequential incubations can already be carried out in the micro-format on MALDI targets and measured directly by mass spectrometry, which greatly simplifies sequence analysis and is of importance for glycoprotein research.

Table 23.11 Example of a parallel enzymatic digestion (enzyme array) of an N-linked oligosaccharide. Source: according to Qian, J. *et al.* (2007) *Anal. Biochem.*, **364,** 8-18.

Enzyme and reagent	Digest number					
	1	2	3	4	5	6
Dried oligosaccharide (ng)	2	2	2	2	2	2
Water (μl)	16	14	12	10	8	6
Reaction buffer, fivefold concentrated (μl)	4	4	4	4	4	4
α(1,3)-Galactosidase (μl)	0	2	2	2	2	2
Sialidase (μl)	0	0	2	2	2	2
β(1,4)-Galactosidase (μl)	0	0	0	2	2	2
β-*N*-Acetylhexosaminidase (μl)	0	0	0	0	2	2
α-Mannosidase (μl)	0	0	0	0	0	2
Total volume (μl)	20	20	20	20	20	20

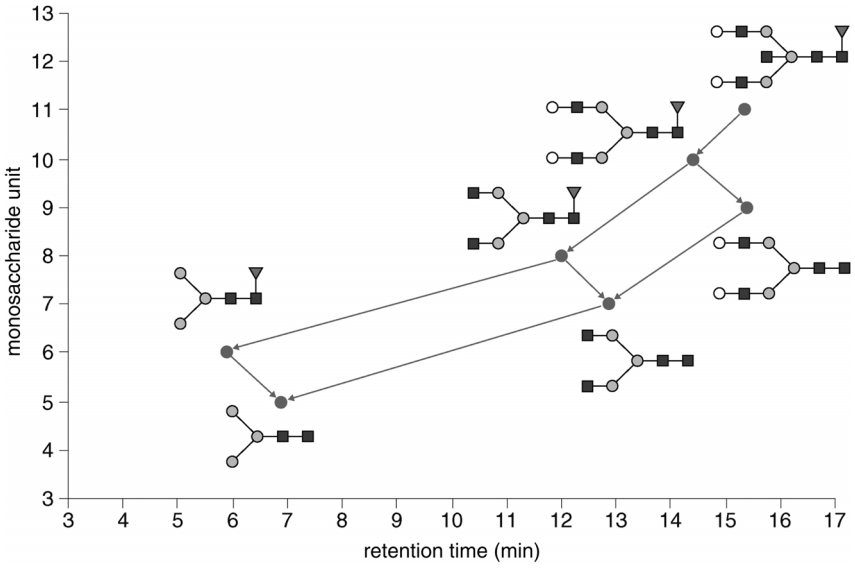


Figure 23.37 Mapping of native *N*-glycans via HPAEC-PAD.

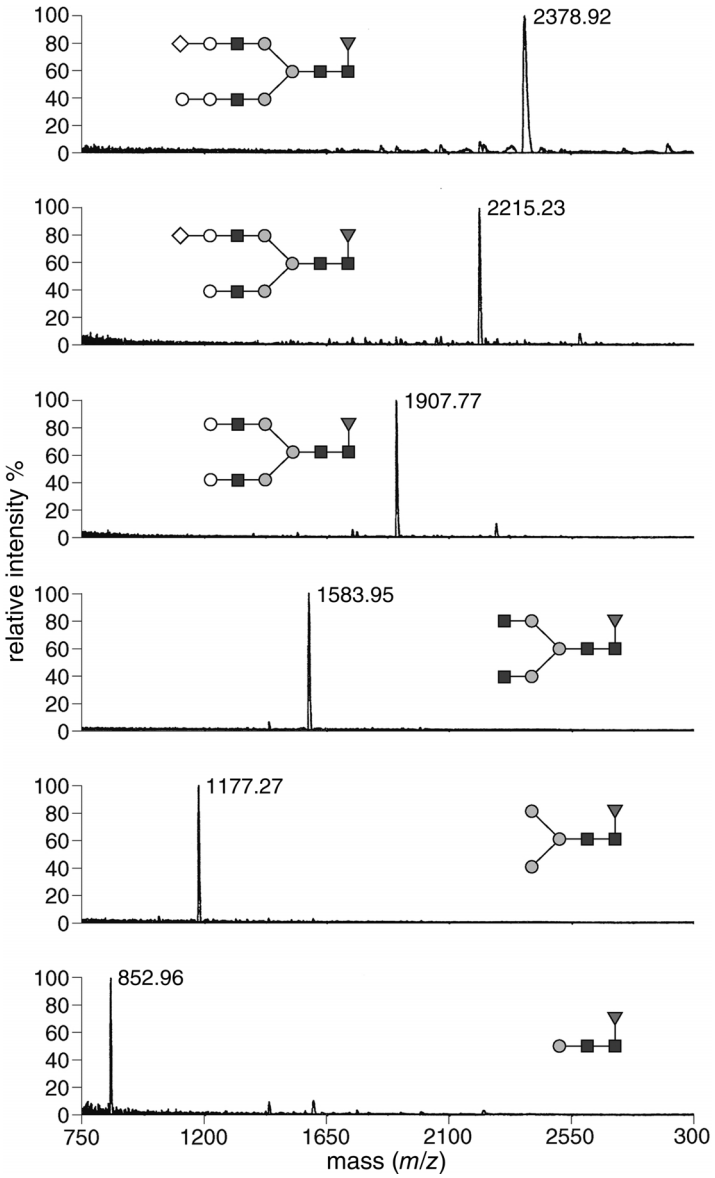


Figure 23.38 MALDI spectra of the sequential enzymatic digestion of the biantennary *N*-glycan G2F-Gal-NGNA carrying proximal fucose, one *N*-glycolyl-neuraminic acid residue and one terminal α(1,3)-linked galactose. Source: according to Qian, J. *et al.* (2007) *Anal. Biochem.*, **364**, 8–18. With permission, Copyright © 2007 Elsevier Inc.

Sequential digestion of an isolated biantennary *N*-glycan of mass *m/z*= 2378.92 (reductaminated by 2-AA) isolated from a monoclonal antibody (cetuximab) provides after each digest a loss of mass that reflects the specificity of the respective enzyme and which is detected in the MALDI spectrum. Thus, digestion with α(1,3)-galactosidase results in a mass loss of 163.69 (– Gal) or a mass loss of 471.15 (– Gal, – NeuGc) if α(1,3)-galactosidase and sialidase are concomitantly present. If in the next step β(1,4)-galactosidase is also present, a further mass loss of 323.82 (– 2Gal) is achieved. Thereafter, the addition of β-*N*-acetylhexosaminidase leads to a mass loss of 406.68 (– 2GlcNAc) and, finally, the addition of α-mannosidase results in a further mass loss of 324.31 (– 2Man).

Mass Spectrometry Both matrix-assisted laser desorption ionization (MALDI-) MS (Section 15.1.1) and electrospray ionization (ESI-) MS (Section 15.1.2) are nowadays among the routine techniques of structural analysis of isolated glycans and glycopeptide fragments (Section 23.3.2). The great advantage of these two techniques is that the glycans no longer must be permethylated (as in the fast atom bombardment (FAB) MS) but can be measured in their native form. Therefore, in recent years FAB-MS has been almost completely displaced by MALDI-MS or ESI-MS.

Since it works in the liquid phase, ESI-MS is particularly well suited for combination with RP-HPLC (Section 10.5) and CE (Section 12.2). The progress made in the miniaturization of HPLC systems (up to micro-LC and nano-LC in packed capillaries) has not only led to a massive increase in the sensitivity and efficiency of this technique but has also improved the routine coupling with ESI-MS technique.

As different mass spectrometric methods have already been described in previous chapters, they are only referred to the investigations covered here (Sections 23.3.2 and 23.3.4). However, it should once again be emphasized that mass spectrometric investigations neither reveal the respective linkage direction and anomery from monosaccharide unit to monosaccharide unit nor at which antenna the respective repeat structure is located. These questions can, however, be answered through ^{1}H NMR spectroscopic investigations, the fundamentals of which are described in the following section.

^{1}H NMR Spectroscopy ^{1}H NMR spectroscopy is the *ultima ratio* of any *N*-glycan structure determination. It allows an unambiguous statement with respect to both the anomeric configuration of each monosaccharide building block and its respective linkage direction towards the neighboring building block.

The basis for this is the fact that in a high-frequency magnetic field the individual protons of the various sugar units may be differentiated with respect to both their respective chemical shift and their respective coupling constant (Section 18.1). This is illustrated below with a few simple examples.

NMR, Chapter 18

The Chemical Shift δ The chemical shift of the H atoms observed in the ^{1}H NMR spectrum differs primarily according to the electromagnetic neighborhood. Therefore, the H atoms of a pyranose ring (e.g., of β-linked galactose) usually experience, as a function of their respective electron-withdrawing neighborhood (by the adjacent O atoms or OH groups), a characteristic shift.

The anomeric proton of a monosaccharide unit is located at a carbon atom (i.e., C1) that is directly linked to two highly electronegative oxygen atoms, that is, the ring oxygen atom on the one hand and the glycosidic oxygen atom on the other hand. Accordingly, a strong (almost double) electron-withdrawing (inductive) effect results at H1. Therefore, the signal of H1 has the largest chemical shift of all ring protons.

Equatorial protons usually come into resonance at lower field than constitutional similar axial protons. This means that H4 of galactose or H2 of mannose (each in an equatorial configuration) exhibit a greater chemical shift than H4 and H2 of the glucose epimer (each in an axial configuration). The situation is similar with the anomeric H-atoms – for example, in glucose: while H1 of α-D-glucopyranose (equatorial configuration) has a chemical shift of $\delta \approx 5.3$ ppm, H1 of the β-anomer (axial configuration) is already in resonance at $\delta \approx 4.7$ ppm. In addition, the anomeric signals differ in their splitting patterns, the so-called coupling constants *J,* which likewise enable determination of the epimeric or anomeric configuration.

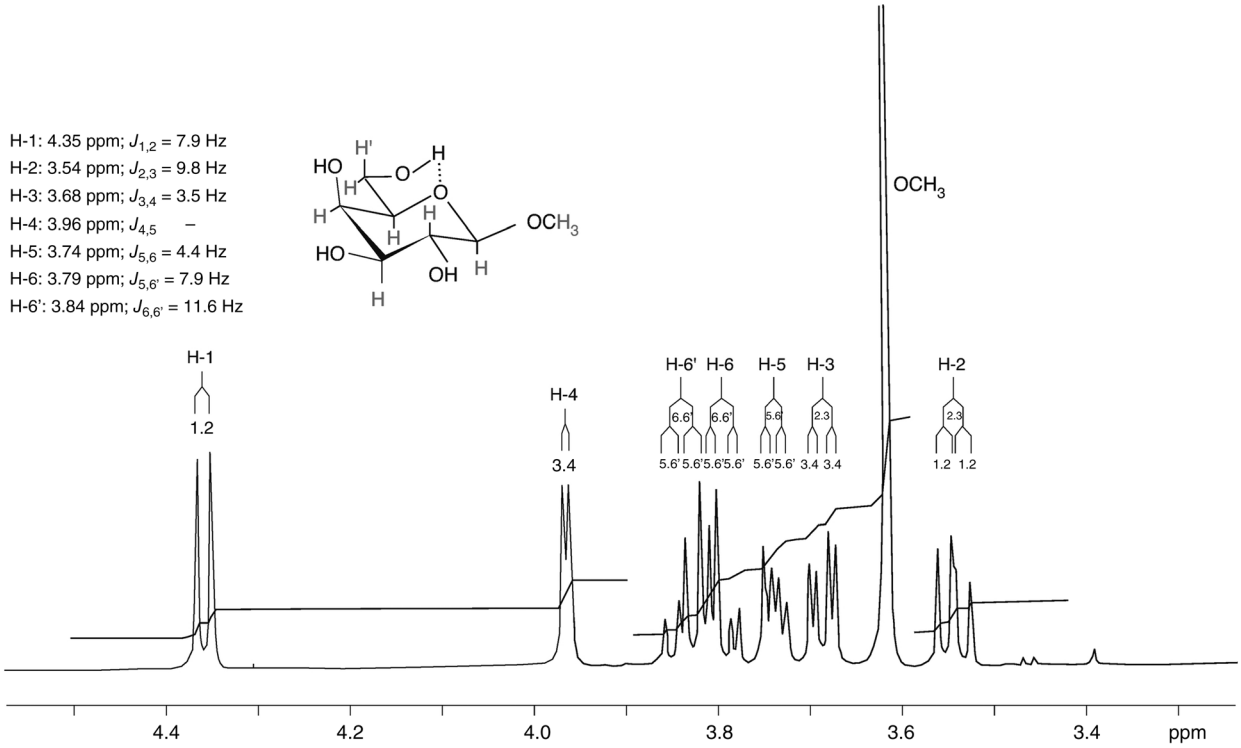


Figure 23.39 ^{1}H NMR spectrum of methyl-β-D-galactopyranoside.

Coupling Constants The coupling constant of a hydrogen atom refers to the width of the respective signal splitting that results from magnetic resonance with an adjacent geminal or vicinal hydrogen atom. This is illustrated in Figure 23.39 using methyl β-D-galactopyranoside as an example.

H1 has only one vicinal neighboring H atom, namely, H2. Thus, the signal of H1 ($\delta = 4.35$ ppm) undergoes (due to magnetic resonance with H2) splitting into a doublet with $J_{1,2} = 7.9$ Hz (large coupling constant due to the antiperiplanar arrangement of H1 and H2).

H2 has two vicinal-antiperiplanar neighboring hydrogen atoms, namely, H1 and H3. As a result, the signal of H2 first splits, through nuclear magnetic resonance with H3 ($J_{2,3} = 9.8$ Hz), into a doublet; and then both signals further split, by coupling with H1 ($J_{1,2} = 8.0$ Hz), in an additional doublet each, so that H2 is detected at $\delta = 3.54$ ppm as a doublet of a doublet.

The situation is similar for H3, which also has two vicinal H-atom neighbors, namely, H2 in antiperiplanar neighborhood ($J_{2,3} = 9.8$ Hz) and H4 in synclinal neighborhood ($J_{3,4} = 3.5$ Hz). Consequently, the signal of H3 ($\delta = 3.68$ ppm) appears as a doublet of a doublet.

Owing to its equatorial configuration H4 has a somewhat larger chemical shift than the axial ring protons at C2, C3, and C5, and comes into resonance at $\delta = 3.96$ ppm. The signal of H4 splits, due to coupling with the adjacent synclinal H3, into a doublet ($J_{3,4} = 3.4$ Hz). Coupling between H4 and H5 is practically not observed.

The signal of H5 ($\delta = 3.74$ ppm) splits first by coupling with H6′ ($J_{5,6} = 7.9$ Hz) into a doublet, and both signals split again into another doublet by coupling with H6 ($J_{5,6} = 4.4$ Hz). The coupling constants indicate a synclinal neighborhood of H5 and H6 and an antiperiplanar neighborhood of H5 and H6′. (This conformational preference of the C5—C6 bond is stabilized by the hydrogen bond between OH6 and the ring-oxygen atom.)

C6 is characterized in that it carries two geminal H-atoms that have a similar chemical shift (H6: $\delta = 3.79$ ppm; H6′: $\delta = 3.84$ ppm). The two protons first come into geminal resonance ($J_{6,6'} = 11.6$ Hz); however, they are still subject to vicinal coupling with H5 ($J_{5,6} = 4.4$ Hz, $J_{5,6'} = 7.9$ Hz) and thus appear as a kind of multiplet.

In β-glucopyranoside, the 4-epimer of β-galactopyranoside, the vicinal neighboring hydrogen atoms are each in an antiperiplanar arrangement ($\Phi = 180°$), and their coupling constants are therefore each at a maximum ($J = 7–10$ Hz).

In contrast to β-galactopyranoside ($\Phi_{1,2} = 180°$), H1 in the corresponding α-anomer takes a vicinal angle to H2 of approximately 60° (synclinal arrangement), showing up in the spectrum of the α-anomer as a relatively small coupling constant ($J_{1,2} = 3–4$ Hz).

Structural Reporter Groups In the ^{1}H NMR spectra of complex *N*-glycans, the peaks of the ring protons coincide in a mountain of overlapping signals between $\delta = 3.5$ and 4.0 ppm, so that the assignment of these protons initially seems impossible. Close by, in the range of $\delta = 4.0–5.5$ ppm and $\delta = 1.5–3.0$ ppm, some clearly separated and therefore characteristic peaks of protons show up, the so-called structural reporter groups (Figure 23.40; box below). They include, between $\delta = 4.0$ and 5.5 ppm, the anomeric protons, which in a β-glycosidic linkage (e.g., β-Gal and β-GlcNAc) reveal a large coupling constant ($J_{1,2} = 7–9$ Hz) and in α-glycosidic bond (e.g., α-Fuc) a small coupling constant ($J_{1,2} = 2–4$ Hz). For mannose, the coupling constant is particularly small ($J_{1,2} = 1–2$ Hz) in both the α- and β-glycosidic linkage, so that the anomeric configuration of mannose cannot be derived from the coupling constant but rather from the chemical shift: the equatorial H1 of the α-mannoside has a higher chemical shift than the axial H1 of the β-mannoside. Moreover, H2 of the mannose residues may be identified in the range $\delta = 4.0–5.5$ ppm. In the region $\delta = 1.5–3.0$, at approximately 2 ppm, the methyl groups of the acetyl groups are detected as large peaks, and next to them, on the right, are the axial and on the left are the equatorial H3 of Neu5Ac, respectively.

Two-dimensional NMR Spectroscopy, Section 18.1.3

As may be seen from the example of the disialo-(α2,3) biantennary *N*-glycan (Figure 23.40), the structural reporter groups provide helpful insight into the nature of the *N*-glycans. Since individual NMR signals are sensitive to pH changes, it is important to perform the measurements under standardized and calibrated conditions (temperature, solvent, pH, reference). However, it is always advisable to interpret the structure obtained by means of ^{1}H NMR data in conjunction with mass spectrometric data, as in the glycan substituents may sometimes be present that do not affect the chemical shift of the structural reporter groups.

If in addition to the structural reporter groups the distinct ring proton signals of an *N*-glycan are to be assigned, you then depend on the more complex 2D NMR experiments. Here the homonuclear coupling experiments according to Hartmann–Hahn (HOHAHA), correlation spectroscopy (COSY), and total correlation spectroscopy (TOCSY) should be mentioned. These experiments allow assignment of the proton signals to distinct spin systems (monosaccharide building blocks). A statement with respect to monosaccharide sequence, linkage direction, torsion angle, or spatial interactions between individual sugar components is finally obtained by 2D nuclear Overhauser enhancement spectroscopy (NOESY). However, 2D NMR spectroscopy of glycans is beyond this introduction to glycoanalysis.

Nuclear Overhauser Effect, Section 18.1.2

Box 23.1 Structural reporter group signals of carbohydrate chains of glycoproteins

Anomeric protons

Amide protons

Man H2

GalNAc-ol H2, H3, H4, and H5

Neu5Ac H3

Fuc H5 and H6

Gal H3 and H4

Protons that are set apart from the mass of peaks due to glycosylation

Protons that are shifted due to substituents such as acyl, sulfate, or phosphate

Protons belonging to substituents such as *O*-methyl, *N*/*O*-acetyl and *N*-glycolyl

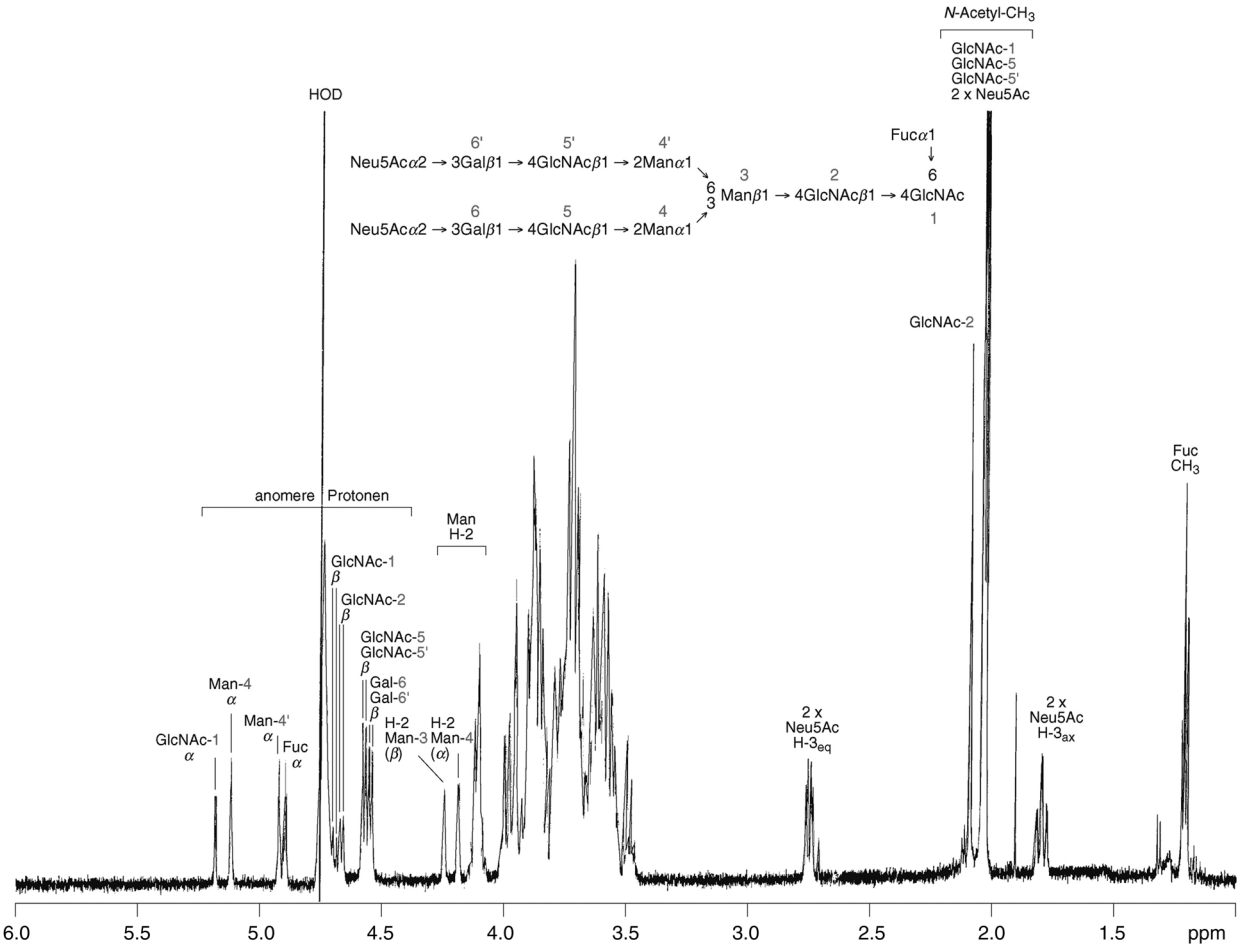


Figure 23.40 600 MHz ^{1}H NMR spectrum of the disialo-(α2,3) biantennary *N*-glycan of rhEPO (in D_2O). The spectrum reveals that the reducing GlcNAc unit (GlcNAc-1) is in the anomeric equilibrium. The signal of the α-anomer appears, with the expected small coupling constant of $J_{1,2\alpha} = 2$–4 Hz, at a higher chemical shift ($\delta = 5.18$ ppm) than the signal of the corresponding β-anomer ($\delta = 4.69$ ppm), which exhibits the large coupling constant ($J_{1,2} = 7$–9 Hz) that is characteristic for the β-anomer. Closely adjacent to that signal is H1 of GlcNAc-2 ($\delta = 4.66$ ppm), in which the ß-glycosidic linkage is again apparent from the large coupling constant ($J_{1,2} = 7$–9 Hz). H1 of the α1,6-linked fucose is visible at $\delta = 4.9$ ppm, with the expected small coupling constant of $J_{1,\ 2\alpha\text{-Fuc}} = 2$–4 Hz. However, owing to the anomeric equilibrium of GlcNAc-1, $H1_{Fuc}$ splits into two overlapping signals of doublets. The two α-linked mannose residues appear, with the small coupling constant characteristic of mannose ($J_{1,2Man} = 1$–2 Hz), at $\delta = 5.11$ ppm (Manα1,3) and $\delta = 4.92$ (Manα1,6). The β-linked mannose is hidden in the massive HOD peak. The two antennary β-GlcNAc residues with $J_{1,2GlcNAc} = 7$–9 Hz can be seen at $\delta = 4.57$ ppm, and close by ($\delta = 4.55$ ppm) there are the signals of the two antennary β-Gal residues ($J_{1,2Gal} = 7$–9 Hz).
In addition to the anomeric protons of the individual monosaccharide units, the following *structural reporter groups* may still be assigned: Of the mannoses: $H2_{Man\beta1,4}$ (4.25 ppm) and $H2_{Man\alpha1,3}$ (4.25 ppm); $H2_{Man\alpha1,6}$ is superimposed by the peak group at 4.1 ppm. These signals (as well as the one of $H1_{Man}$) are characterized by their small coupling constants. The equatorially oriented protons $H3_{eq}$ of Neu5Ac enter into resonance at 2.76 ppm and the axial protons $H3_{ax}$ at 1.80 ppm. This $H3_{ax}$ exhibits (in addition to the coupling to $H3_{eq}$) the expected large coupling to adjacent antiperiplanar H4 ($J_{3ax,4} = 7$–9 Hz), while the coupling of the equatorial $H3_{eq}$ with H4 is small because of the synclinal arrangement to H4 ($J_{3eq,4} = 2$–4 Hz). Likewise, CH_3-6 of fucose is clearly exposed as a *reporter group* ($\delta = 1.2$ ppm). Interestingly, the anomeric equilibrium of GlcNAc-1 may still be recognized at the attached fucose: CH_3-6_{Fuc} does not occur as a doublet (due to coupling with $H5_{Fuc}$) but as a kind of triplet (due to superposition of two closely adjacent doublets). The *N*-acetyl peaks of GlcNAc and Neu5Ac are found at $\delta = 2.04$, with the NAc-peak of GlcNAc-2 slightly separated ($\delta = 2.09$ ppm). Source: courtesy of M. Nimtz, Braunschweig.

Nuclear Overhauser Effect (NOE) As we have seen in the previous section, the dipole moments of neighboring protons of the pyranose ring are in magnetic interaction, which is manifest in the ^{1}H NMR spectrum via a characteristic (dependent on the particular torsion angle) signal splitting (coupling constant). If the magnetic interaction of two adjacent protons is now supported by irradiating perpendicularly to the applied high-magnetic field of the NMR spectrometer the magnetic resonance frequency of one of the two protons, the magnetic coupling of this proton to the neighboring proton is amplified (homonuclear NOE). As a consequence, the signal of the neighboring proton is enhanced due to dipole–dipole interaction. Since this signal increase will change with the distance between the two protons in a ratio of r^{-6} (r is the distance between the

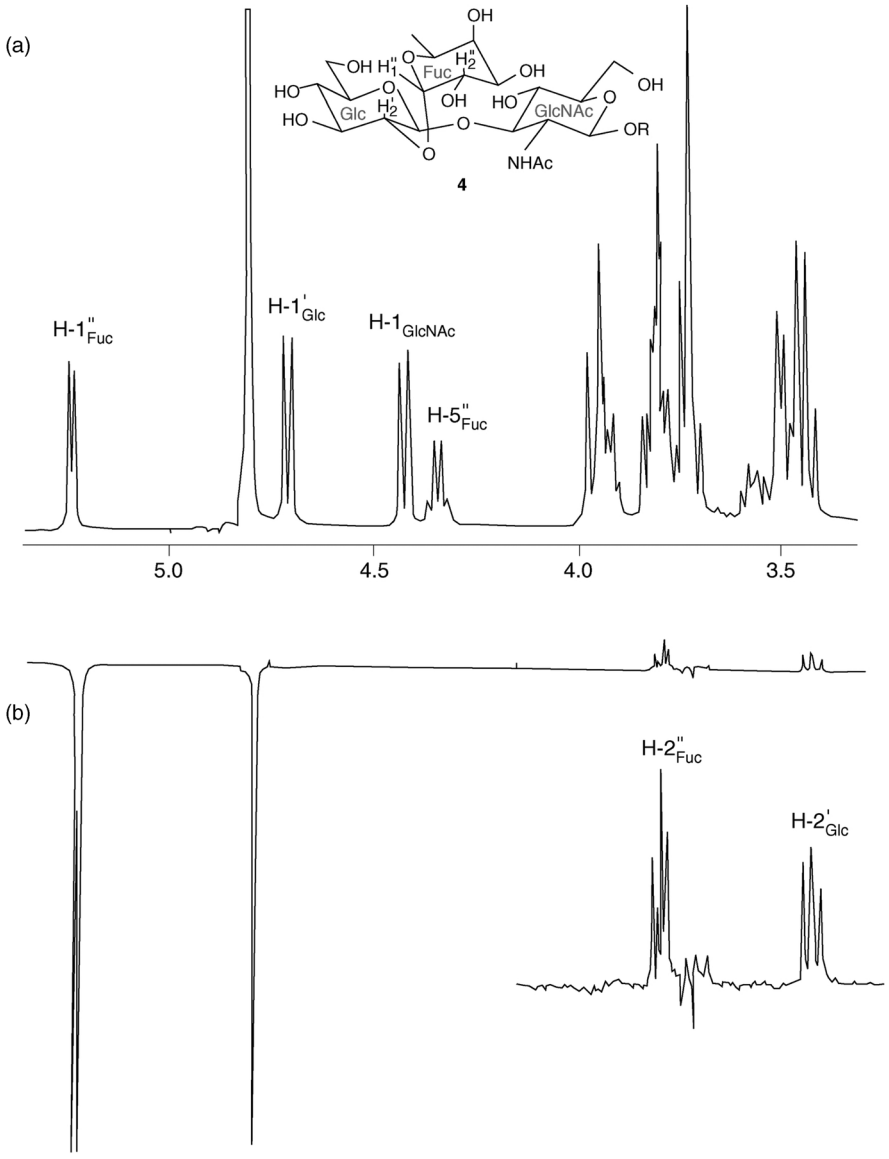


Figure 23.41 Nuclear Overhauser enhancement experiment with the trisaccharide L-Fucα1,2-D-Glcβ1,3-D-GlcNAc (400 MHz ^{1}H NMR spectrum in D_2O). (a) Normal spectrum, (b) difference spectrum upon irradiation of the resonance frequency of $H1''_{Fuc}$. The ^{1}H NMR spectrum of the trisaccharide first discloses the three anomeric protons, which are quite easily recognized by their high chemical shifts. In particular, $H1''_{Fuc}$ of the α-glycosidically linked fucose residue is clearly assigned at $\delta = 5.3$ ppm, due to its small coupling constant of $J_{1,2Fuc} = 4$ Hz; likewise, $H5''_{Fuc}$ may easily be recognized at $\delta = 4.35$ ppm through the quartet characteristic for fucose. The two prominent anomeric signals at $\delta = 4.7$ and $\delta = 4.4$ ppm are each recognized, due to the large coupling constant of $J_{1,2} = 8$ Hz, as H1 of the two β-glycosidically linked Glc and GlcNAc units. The signals of the remaining protons are crowded and superimposed in peak groups in the range 3.4–4.0 ppm, so here a peak assignment seems impossible at first glance.
If in the NOE experiment you now irradiate the resonance frequency of $H1''_{Fuc}$, the resonance saturation of $H1''_{Fuc}$ is transferred, by dipole–dipole interaction, to those protons that enter into nuclear magnetic resonance coupling with $H1''_{Fuc}$; the signals of these protons in the ^{1}H NMR spectrum are therefore increased. These signals are now identified in the Fourier-transformed difference spectrum as clearly attributable peaks. In fact, the difference spectrum reveals $H2''_{Fuc}$ (which is in vicinal neighborhood to proton $H1''_{Fuc}$) as main signal at $\delta = 3.8$ ppm. In addition, the signal of $H2'_{Glc}$ is revealed at $\delta = 3.46$ ppm. The coupling constants of $H2''_{Fuc}$ demonstrate, on the one hand, the synclinal neighborhood to $H1''_{Fuc}$ ($\Phi_{1,2Fuc} = 60° \rightarrow J_{1,2Fuc} = 4$ Hz) and, on the other hand, the antiperiplanar proximity to $H3''_{Fuc}$ ($\Phi_{2,3Fuc} = 180° \rightarrow J_{2,3Fuc} = 8$ Hz). In addition, the coupling constants of $H2'_{Glc}$ reveal, on the one hand, the antiperiplanar proximity to $H1'_{Glc}$ ($\Phi_{1,2Glc} = 180° \rightarrow J_{1,2Glc} = 8$ Hz) and, on the other hand, the antiperiplanar proximity to $H3'_{Glc}$ ($\Phi_{2,3Glc} = 180° \rightarrow J_{2,3Glc} = 8$ Hz). The two NOE effects are of comparable size, namely, 8% for $H2'_{Glc}$ and 7% $H2''_{Fuc}$, which means that the distance from $H1''_{Fuc}$ to its vicinal neighbor $H2''_{Fuc}$ is comparable with its distance to $H2'_{Glc}$ of the neighboring glucose block. This proximity is expressed in the structural formula shown in (a). The NOE effects measured via ^{1}H-NMR spectra thus enable a direct confirmation of the preferred conformation of the glycosidic bond as predicted on the basis of the exo-anomeric effect. A memory aid is that the hydrogen atoms adjacent to the glycosidic oxygen atom (i.e., $H1''_{Fuc}$ and $H2'_{Glc}$ as well as $H1'_{Glc}$ and $H3_{GlcNAc}$), in proper projective representation, always must take a synperiplanar (so to speak, parallel) arrangement (Section 23.1.5). Source: Hindsgaul, O. (1980) Ph.D. thesis, the University of Alberta, Edmonton.

irradiated proton and the proton for which the signal increase is observed), such measurements are highly sensitive with respect to the smallest stereochemical changes. If the two protons are located on different monosaccharide units of the oligosaccharide, even a minimal conformational change will alter the NOEs, provided the spatial arrangement of the two protons is altered. As a 10–20% increase of signal intensity can hardly be seen in the NMR spectrum, the signal increase may be determined using the following trick: the two NMR spectra are subtracted from each other (this is done by the NMR software). Thus, all peaks of the NMR spectrum that have remained unchanged will cancel each other, and a smooth baseline will be observed. Only where an increased dipole–dipole interaction is observed in the spectrum by a peak increase will a non-zero residual peak be detected after subtraction of the two spectra – exactly at a chemical shift that is characteristic for the adjacent proton. The relative NOE values of two or more protons within a molecule thereby provide a direct measure of their spatial distance. This fact, which is somewhat difficult to understand, is illustrated using the trisaccharide shown in Figure 23.41 by way of example.

Spatial Interaction of Sugar Residues As already mentioned, the glycosidic bonds of oligosaccharides exhibit spatial preferences with respect to linkage direction and conformation due to the exo-anomeric effect (Section 23.1.5). In branched carbohydrate determinants a given hydrogen atom may spatially be in the neighborhood of strongly electronegative oxygen atoms. In such a case an inductive (spatial) interaction of this proton is to be expected, which in the ^{1}H NMR spectrum results in an increased chemical shift. This is illustrated in Figure 23.42.

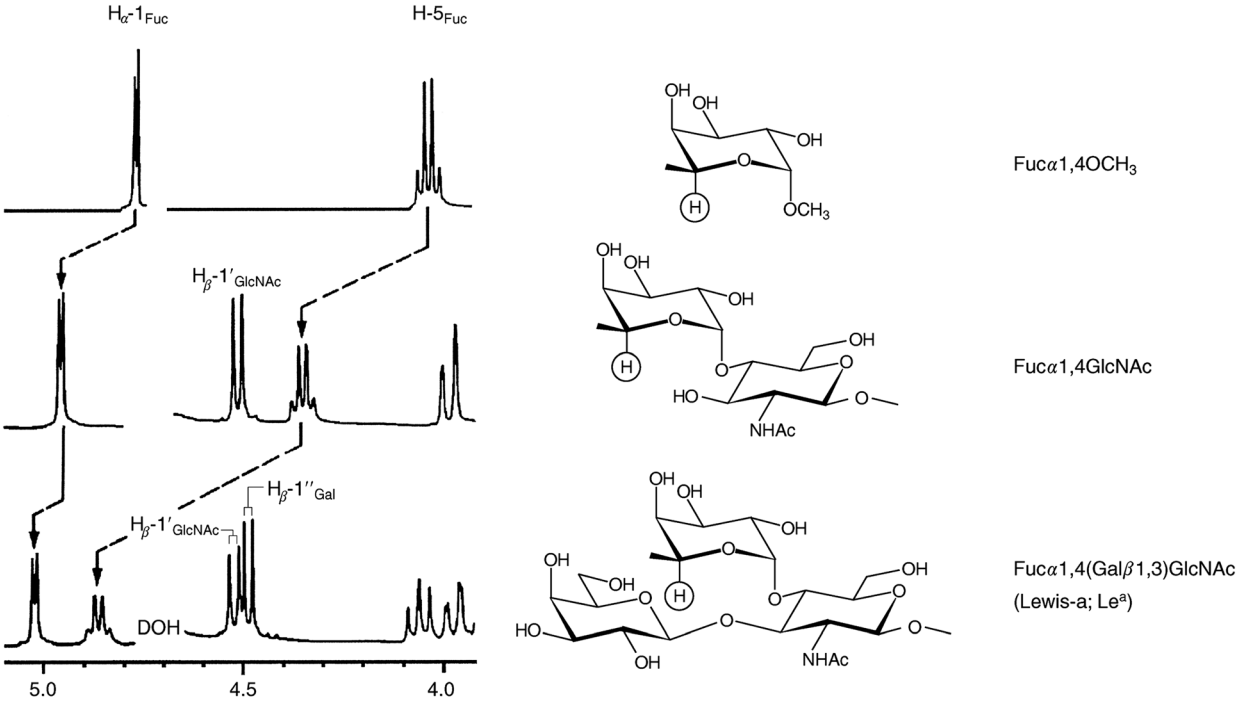


Figure 23.42 Detection of spatial interaction in the Lewis-a blood group determinant by means of ^{1}H NMR spectroscopy. The ^{1}H NMR spectrum of monosaccharide building block methyl-α-L-fucopyranoside exhibits the anomeric proton at the high chemical shift of $\delta = 4.75$ ppm characteristic for α-anomers, and the coupling constant to the adjacent vicinal proton H2 ($J_{1,2} = 4$ Hz) corresponds to the splitting expected for a torsion angle of 60°. At the same time, the ^{1}H NMR spectrum displays H5 at a prominent place ($\delta = 4.05$) as a quartet characteristic for 6-deoxy sugars, the peak splitting of which is due to the coupling with H4 and the 6-methyl group. If now the L-fucopyranosyl block is in α-glycosidic linkage to OH4 of a β-D-GlcNAc residue, the ^{1}H-NMR spectrum reveals for both $H1_{Fuc}$ as well as for $H5_{Fuc}$ a significant downfield shift to $\delta = 4.95$ ppm (for $H1_{Fuc}$) and $\delta = 4.35$ ppm (for $H5_{Fuc}$). At the same time, the spectrum displays the anomeric proton of the β-D-GlcNAc residue as a doublet at a chemical shift of $\delta = 4.5$ ppm – with the high coupling constant of $J_{1,2GlcNAc} = 8$ Hz characteristic for a β-glycoside (because of the antiperiplanar arrangement of H1 and H2). As can be seen from the figure, the shift of the signals of $H1_{Fuc}$ and $H5_{Fuc}$ may be explained by the added spatial neighborhood of OH3 of the GlcNAc-block – in the illustrated conformation $OH3_{GlcNAc}$ induces a spatial inductive effect on $H1_{Fuc}$ and $H5_{Fuc}$ that is manifest from the shifts observed in the ^{1}H NMR spectrum. If $OH3_{GlcNAc}$ is now associated with a β-galactosyl residue, as is the case in the trisaccharide determinant of human blood group Lewis-a, especially $H5_{Fuc}$ experiences an additional downfield shift to $\delta = 4.85$ ppm. This additional shift is explained by the fact that now $H5_{Fuc}$ is also in spatial proximity to the ring oxygen atom of the newly added β-Gal-block. At the same time the ^{1}H NMR spectrum of the trisaccharide exhibits (in close proximity to the signal of $H1_{GlcNAc}$) the newly added signal of $H1_{Gal}$, which, due to the β-glycosidic linkage, shows the expected large coupling constant of $J_{1,2Gal} = 8$ Hz (like $H1_{GlcNAc}$). Source: according to Lemieux, R.U. (1990) *Explorations with Sugars. How Sweet it Was*, American Chemical Society, Washington D.C.

The example shows that the ^{1}H NMR spectra of carbohydrate determinants may reveal important information about the configurational arrangement of adjacent vicinal H atoms as well as the conformational and spatial neighborhood of monosaccharide units or their H atoms and OH groups. Therefore, ^{1}H NMR spectroscopy is now regarded as one of the most important analytical techniques of carbohydrate chemistry.

23.4 Genome, Proteome, Glycome

Following elucidation of the human genome (2001), the proteome, the totality of all proteins existing in a cell or tissue under defined conditions, has come into the focus of biomedical research. The functional characterization of the proteins encoded by the genes and expressed in living organisms promises direct intervention into the molecular causes of diseases. Therefore, protein glycosylation, which alters the properties and functions of proteins, is of particular importance in that regard: Glycome research examines the total set of all carbohydrate structures of an organism, whereas glycoproteome research analyzes the protein-bound totality of all glycans of an organism.

Glycome or glycoproteome research has the following goals:

- on the genome level: to identify all the genes relevant for the glycans of an organism;
- on the proteome level: to locate on the identified glycoproteins the currently occupied glycosylation sites;
- on the glycome level: to analyze and functionally assign the respective (organism-, time-, and locus- or organ-dependent as well as disease-related) glycosylation status of the glycoproteins.

The special challenge of glycoproteome research is that the target structures, the glycoproteins (to which we restrict ourselves in this chapter) and their glycans, are not direct gene products but, instead, the carbohydrate side chains are only post-translationally built and rebuilt via special enzymes (glycosyltransferases that form glycosidic bonds in a stereo- and regiospecific manner). These syntheses within the organisms are not always complete, so that glycoproteins usually exhibit mixtures of heterogeneous glycan structures (cf. microheterogeneity, glycoforms, Section 23.2).

In addition, in organisms there are often different glycosyltransferases that compete for a common glycosyl acceptor. Thereby, the glycostructures usually become more diverse and complex. Thus, it is virtually impossible to deduce the glycan structures directly from the genome information – in contrast to the protein structures, which represent direct gene products.

Genome sequencing has revealed that in organisms there are numerous genes whose respective function is completely unknown. If such genes should have a meaning for glycosylation, glycome or glycoproteome research could provide an approach to elucidate the function of these genes.

But while proteomics researchers already now use robot-assisted high-throughput screening to analyze fully automatically micro-amounts of biomedical relevant samples from gradient-focused high-resolution 2D-gels (in conjunction with MALDI-MS or ESI-MS) within a very short time, similar approaches in glycome or glycoproteome research only just appearing.

A recently published research approach of Australian and American scientists uses a robot-assisted microdispension system and consists of:

1. PNGase F digestion of the (glyco)protein patterns blotted onto PVDF membranes from 2D electrophoresis gels
2. Extraction with water of the liberated *N*-glycans, followed by analysis via LC-ESI-MS (Section 15.2.6).
3. Then the *O*-glycosidically linked sugar chains are released from the protein spots by chemical means (reducing β-elimination) and likewise extracted with water and analyzed via LC-ESI-MS.
4. Finally, the deglycosylated protein spots remaining on the membrane are digested with trypsin and the peptide fragments analyzed and sequenced via MALDI-TOF-MS (Section 15.1.5).

5. The amino acid sequences obtained are compared with existing genomic or proteomic databases to infer the associated structural genes or proteins.

In an alternative approach, Japanese researchers use lectin-affinity chromatography (Section 23.3.1) – instead of 2D electrophoresis – that can specifically recognize entire glycan groups (e.g., the *high-mannose*, complex, or hybrid type; Figure 23.10).

In this glyco-catch method, the cells or tissues to be tested are first homogenized and then centrifuged and treated as follows. The target glycoproteins are:

1. selectively enriched via lectin-affinity chromatography – directly from the soluble supernatant (or indirectly from the insoluble sediment – after treatment with 1% Triton X-100 to solubilize the membrane (glyco)proteins);
2. comprehensively digested by means of a highly specific protease (preferably lysyl endopeptidase, which is fully active even in the presence of 4 M urea and 0.1% SDS and likewise destroys undesirable hydrolytic enzymes such as glycosidases or proteases);
3. isolated via a lectin-affinity column of the same specification as in step 1;
4. separated via RP-HPLC (Section 10.6);
5. if sufficient material is available, the relevant HPLC fractions are subjected to amino acid sequencing (Edman degradation, Section 14.1);
6. the peptide sequences obtained are finally compared with existing genomic or proteomic databases to identify the corresponding structural genes or proteins.

If the specificity of the used lectin is sufficiently known, the type of glycosylation may in principle already be predicted from the lectin type. The actual carbohydrate structure side-chain assignment, however, must rely on a combination of the various analytical techniques presented in this chapter, in order to finally confirm the identified glycan structures. In this respect, mass spectrometric methods are of particular importance.

23.5 Final Considerations

As we have seen, structural characterization of the glycans of a glycoprotein is a very complex, technical, and time-consuming task that needs highly trained staff. This requires an arsenal of different analytical techniques that are mostly used on the intact glycoprotein, partly on the proteolytic glycopeptide fragments with individual glycosylation sites, and finally on the released native or derivatized glycan pool or the isolated individual structures.

Today, the aim is to develop methods and techniques that make it possible to replace the very expensive and time-consuming and costly analysis with simpler, faster, and cheaper standardized procedures. Here, the coupling of the orthogonal analytical techniques of HPLC and mass spectrometry (MALDI, ESI) and the development of glyco-databases that enable routine online-analysis of the glycans of glycoproteins are of particular importance.

The combined HPLC-MS(MS) methods – with direct glycopeptide or glycan measurement – enable us, on the one hand, to determine the molecular weights of the glycans (from the MS data of the peaks separated via HPLC) and, on the other hand, to draw conclusions about the respective glycan substructures (from the MS/MS fragmentation patterns). The actual glycan structures are finally selected from the structural isomers allowed by mass spectrometry by comparing the precisely determined HPLC retention times with a database or with standard structures. Thus, HPLC-MS(MS) coupling in connection with a valid mapping database opens up new perspectives for online determination of glycan structures – with a direct stimulatory feedback for ongoing research projects and massive time savings.

In the pharmaceutical industry, most of the analytical methods or techniques presented herein require their reliability to be secured by extensive validation under GMP (good manufacturing practice) and GLP (good laboratory practice) conditions, that is, documentation of their specificity, precision, reproducibility, accuracy, and linearity – regardless of test laboratory and laboratory personnel. However, the biological safety and therapeutic efficacy of bioengineered glycoproteins can ultimately only be determined in extensive clinical trials.

Further Reading

Fukuda, M. and Kobata, A. (ed.) (1993) *Glycobiology. A Practical Approach*, Oxford University Press, Oxford.

Gabius, H.-J. and Gabius, S. (eds) (1996) *Glycosciences: Status and Practice*, Chapman & Hall, Weinheim.

Geyer, H., and Geyer, R. (2006) Strategies for analysis of glycoprotein glycosylation. *Biochem. Biophys. Acta*, **1764**, 1853–1869.

Higgins, E. (2010) Carbohydrate analysis throughout the development of a protein therapeutic. *Glycoconj. J.*, **27**, 211–225.

Hirabayashi, J. (2004) Lectin-based structural glycomics: glycoproteomics and glycan profiling. *Glycoconj. J.*, **21**, 35–40.

Hirabayashi, J. *et al*. (2002) Glyco-catch method: a lectin affinity technique for glycoproteomics. *J. Biomol. Tech.*, **13**, 205–218.

Kawasaki, N. *et al*. (2000) Application of liquid chromatography/mass spectrometry and liquid chromatography with tandem mass spectrometry to the analysis of the site-specific carbohydrate heterogeneity in erythropoietin. *Anal. Biochem.*, **285**, 82–91.

Lemieux, R.U. (1990) *Explorations with sugars. How sweet it was*. American Chemical Society, Washington D.C.

Lipton, M. and Pasă-Tolic, L. (eds) (2009) *Mass Spectrometry of Proteins and Peptides*, vol. 492, Humana Press.

Nimtz, M. *et al*. (1993) Structures of sialylated oligosaccharides of human erythropoietin expressed in recombinant BHK-21 cells. *Eur. J. Biochem.*, **213**, 39–56.

Qian, J. *et al*. (2007) Structural characterization of N-linked oligosaccharides on monoclonal antibody cetuximab by the combination of orthogonal matrix-assisted laser desorption/ionization hybrid quadrupole-quadrupole time-of-flight tandem mass spectrometry and sequential enzymatic digestion. *Anal. Biochem.* **364**, 8–18.

Rassi, Z.El (ed.) (1995) *Carbohydrate Analysis. High Performance Liquid Chromatography and Capillary Electrophoresis*, Journal of Chromatography Library, vol. 58, Elsevier, Amsterdam.

Sloane, A.J. *et al*. (2002) High throughput peptide mass fingerprinting and protein macroarray analysis using chemical printing strategies. *Mol. Cell Proteomics*, **1**, 490–499.

Wilson, N.L. *et al*. (2002) Sequential analysis of N- and O-linked glycosylation of 2D-PAGE separated glycoproteins. *J. Proteome Res.*, **1**, 521–529.

Lipid Analysis

24

Harmut Kühn[1] and Klaus van Leyen[2]

[1]Institute of Biochemistry, Charité-Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-University zu Berlin, and Berlin Institute of Health, Charitéplatz 1, 10117 Berlin, Germany
[2]Neuroprotection Research Laboratory, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, 149 13th St. R. 2401, Charlestown, MA 02129, USA

Along with proteins, nucleic acids, and carbohydrates, lipids form the fourth major class of natural substances. The common characteristic of all lipids is their low solubility in water (high degree of hydrophobicity) and their high solubility in organic solvents. The biological functions of lipids are just as variable as their structures and can be summarized as follows:

1. Energy storage and heat insulation: In many organisms, fats and oils represent the major long-term energy storage substances. In animals, specifically triacyl glycerols are ideal storage lipids and heat insulators. Deposited in white and brown fatty tissue, triacyl glycerols can be mobilized when needed (hibernation, fasting).
2. Structural scaffolds and barrier function: Due to their low solubility in water and their amphiphilic character, some lipid classes can aggregate as micelles in an aqueous environment. The ability of phospholipids to form lipid bilayers in aqueous solutions constitutes the structural basis for the formation of biomembranes and, thus, for cell compartmentalization.
3. Lipids as signal transducers: In plants and animals lipids function as hormones (e.g., steroid hormones, eicosanoids, jasmonic acid), as intracellular messengers (e.g., diacyl glycerols, sphingosine), and as components of electron transport systems (ubiquinone). They transport information.
4. Lipidic vitamins: As vitamins, lipids play major roles in the visual system (vitamin A), in the coagulation cascade (vitamin K), and in regulation of extracellular calcium homeostasis (vitamin D). Vitamin E and other lipophilic antioxidants function as radical scavengers, and protect biomembranes and lipoproteins against oxidative damage.
5. Defense mechanisms: Many lower organisms and fungi produce complex lipid derivatives with antibacterial properties (antibiotics). Some of them are synthesized via the polyketide pathway and typically block processes important for bacterial replication, such as gene transcription and/or proteins synthesis. Typical examples are the antibiotics tetracycline and erythromycin.

24.1 Structure and Classification of Lipids

Lipids can be classified either structurally or by their functions. Structure-based classification systems rely on the fact that lipids are composed of smaller structural units that give the lipid class its name. Following the recommendation of an international panel of experts in lipid chemistry a comprehensive classification system was conceived in 2005 (and later refined in 2009), which classifies all biologically relevant lipids into eight different groups:

1. fatty acids and their derivatives (FA)
2. glycerolipids (GC)

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

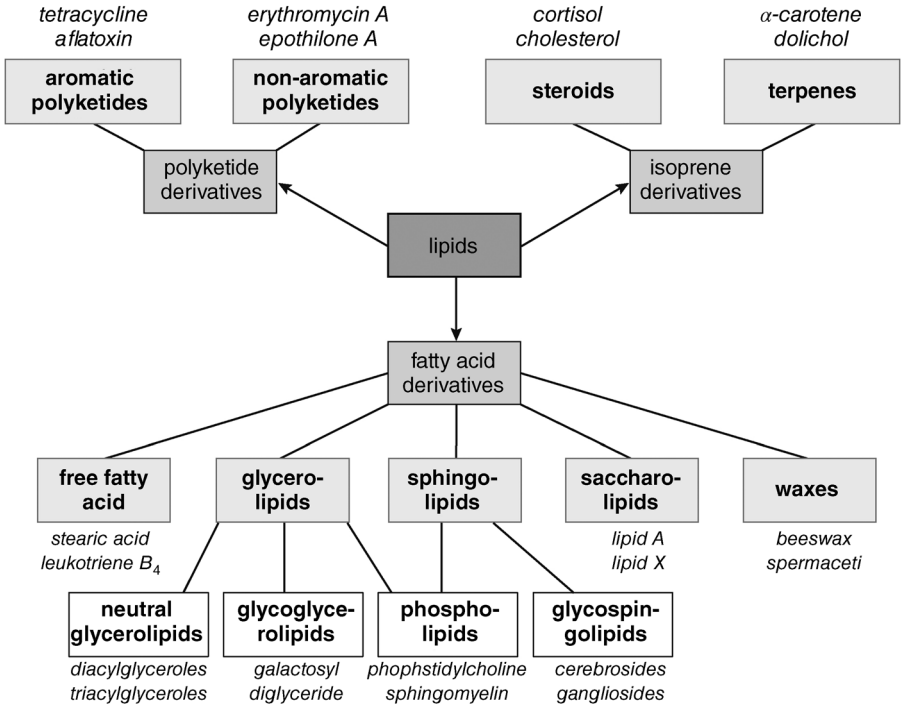


Figure 24.1 Classification of biologically relevant lipids. In this classification system lipids are divided into three major classes – fatty acid derivatives, isoprenoids, and polyketides. Each of these major classes consists of several subclasses, which can be subdivided. Important members of each class are shown in italics.

3. glycerophospholipids (GP)
4. sphingolipids (SP)
5. sterols (ST)
6. prenols (PR)
7. glycolipids (GL)
8. polyketides (PK)

This comprehensive classification system is very useful and allows systematic classification of all naturally occurring lipids. However, some lipid molecules can be assigned to multiple lipid classes and thus the system is not unequivocal. For example, by definition the glycerophospholipids are a subset of glycerolipids and thus may be included in both subclasses. Figure 24.1 suggests an alternative lipid classification scheme, which is based on the idea that all lipids can be classified in three main classes (fatty acid derivatives, isoprenoids, and polyketides). These main lipid classes can be further subdivided into subclasses.

Fatty acid derivatives are the most abundant lipids in the human body. They contain fatty acids, which are medium- to long-chain univalent carboxylic acids. They can react with alcohols or amines to generate ester lipids, ether lipids, and amino lipids. Their chemical characteristics are mainly determined by the fatty acid moiety. Table 24.3 below summarizes some of the biologically most relevant fatty acids. This main lipid class also includes native and modified free fatty acids, such as eicosanoids.

Like fatty acids, isoprenoids are synthesized in biological systems from acetyl-CoA. While fatty acids are synthesized via the intermediate formation of malonyl-CoA, isoprenoid biosynthesis involves the production of activated isoprene in the shape of isopentenyl pyrophosphate or dimethylallyl pyrophosphate. By coupling of several activated isoprene units either linear isoprenoids (e.g., terpenes) or, following cyclization, cyclic isoprenoids (e.g., sterols) are generated.

Polyketides form a large family of lipids that are very heterogeneous in terms of their chemical structures and their biological relevance. They all have in common their biosynthetic origin via the polyketide pathway, which involves the joining of short-chain coenzyme A-activated acyl groups. The stepwise chain extension resembles the mechanism of fatty acid biosynthesis, but the coupling reaction is not accompanied by complete reduction of the carbonyl group. During biosynthesis each acyl unit can be chemically modified, leading to a multitude of products that often leaves the original polyketone structure unrecognizable.

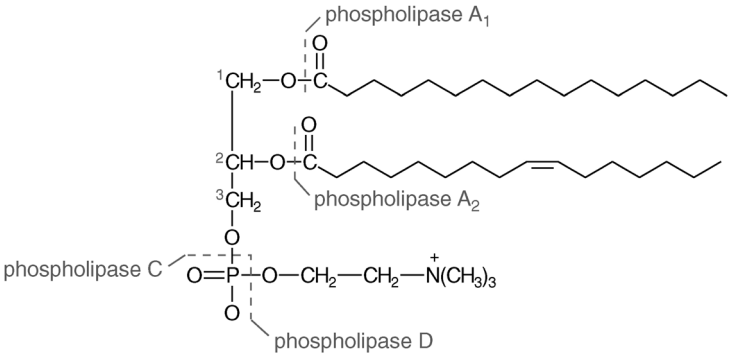


Figure 24.2 Phospholipases of varying specificity hydrolyze phospholipids at predetermined positions, thereby enabling determination of the position at which a hydrolyzed fatty acid is bound.

Modifications include, among others, reductions, cyclizations, methylations, and oxidations. One example of a simple biologically active polyketide is 6-methylsalicylic acid. Among the more complex polyketides are plant dyes such as quercetin, the mycotoxin zearalenone, the macrolide antibiotic erythromycin, the antifungal amphotericin B, and the immune suppressor rapamycin.

Biological membranes and lipoproteins are lipid–protein assemblies, whose function and chemical properties are largely determined by the nature of their lipid components. For example, biomembranes with a high percentage of cholesterol, or a low percentage of polyunsaturated fatty acids, exhibit reduced fluidity. If this occurs in the plasma membrane of circulating blood cells, the cells are more susceptible to changes in osmotic pressure and shear stress, resulting in a reduced lifespan of the cell. A change in the lipid composition of membranes can impact both signal transduction processes, and the activity of membrane-bound complexes. Quantitative analysis of changes in the lipid compartment of biomembranes is thus of high importance. Lipid hormones and intracellular second messengers play key roles in both intracellular and extracellular signaling. In several endocrine diseases an altered synthesis of lipid hormones is the cause of complex pathology. For diagnosis of these types of metabolic diseases accurately quantifiable analytical methods are required that can be used for routine quantification.

A general procedure for lipid analysis is carried out in three steps:

1. Preparation and purification of lipids from biological sources. In addition to lipid extraction this step commonly includes several chromatography procedures. For a large number of medically relevant lipids quantification methods have been worked out that do not require pre-purification.
2. Analysis of the isolated lipid by elemental analysis, UV or IR spectroscopy, mass spectrometry or NMR. In recent years electrospray ionization mass spectrometry (ESI-MS) has been particularly successful, either as single ESI-MS or as tandem ESI-MS/MS. This analytical method is the basis of current lipidome analysis (Section 24.6).
3. Lipid fragmentation and analysis of the fragmentation products. Fragmentation can be induced by enzymatic or non-enzymatic reactions. Alkaline hydrolysis non-selectively cleaves all ester bonds, whereas acid hydrolysis predominantly cleaves amide bonds of sphingolipids and enol ether bonds of the plasmalogens. Selective enzymes, for example, triacyl glycerol lipases and cholesterol esterases, cleave only their respective substrate lipids. Various types of phospholipases cleave phospholipids at specific positions (Figure 24.2). A common fragmentation method for esters lipids is trans-esterification (e.g., trans-methylation), where complex fatty acid esters are converted into the corresponding fatty acid methyl esters, which are suitable for subsequent gas chromatography.

24.2 Extraction of Lipids from Biological Sources

Besides the traditional liquid phase extraction using organic solvents, in recent years solid phase extraction with hydrophobic carrier materials has gained in importance. To enable an exact quantification of the extraction procedure and to judge the recovery of extracted lipids, external

quantification standards are added to the biological material. These standards should be structurally related to the lipids that will be analyzed. However, they must be distinct enough to allow separate quantification. For fatty acid analysis, esters of non-physiological fatty acids may be used as external standards. Prostaglandin B_2 is commonly used as external standard in eicosanoid analysis (Section 24.4.5). If the lipids are analyzed by mass spectrometry, deuterated or ^{18}O-labeled compounds can be added as external standards.

24.2.1 Liquid Phase Extraction

To extract the total lipid fraction from cells or tissues mechanical homogenization is usually the first step. If tissue homogenization is performed in water or buffer, lipases might be activated. This problem can be circumvented by addition of a polar solvent, or by cooling the homogenate in an ice bath. Mixtures of organic solvents that are non-miscible with water, such as hexane, diethyl ether, chloroform, dichloromethane, or ethyl acetate, are frequently used as extraction solvents. Using a suitable mix of polar and nonpolar organic solvents, particular lipid classes can be extracted more or less selectively.

Besides the activation of lipases, lipid peroxidation is the second major challenge. Many lipids contain unsaturated fatty acids and thus are quite susceptible to radical-mediated oxidation reactions. Lipid extractions should therefore be carried out at low temperatures (ice bath or cold room) under an inert gas atmosphere, such as argon. Ideally all solvents should be purged of oxygen prior to use by flushing with argon gas.

The most common method for lipid extraction is known as Folch extraction. Here, biological tissues are homogenized in a 2 : 1 mixture of chloroform and methanol, typically using about 20 ml of Folch mix for 1 g of wet tissue. Precipitated proteins and nucleic acids are filtered off, and may be re-extracted. If the extracted tissue does not have a high water content, the extract is homogeneous at this stage. The lipid extract is then washed with water or a salt solution (e.g., 1 M NaCl) in a separation funnel until phase separation is achieved. The organic phase, which depending on the solvents used can be either the upper or the lower phase, contains the extracted lipids and can be used for further analysis. Neutral lipids, phospholipids, most sphingolipids, and also lysophosphatidic acids are nearly quantitatively extracted using this method. Complex glycolipids with high carbohydrate content are only partially recovered in the organic phase. They can be isolated by solid phase extraction from the aqueous phase of the Folch extract.

As an alternative to the Folch extraction, lipids can be isolated by the method of Bligh and Dyer. This method is particularly useful when large amounts of biological material are to be extracted, and quantitative extraction is not required. In the typical Bligh–Dyer extraction protocol 2.5 ml methanol and 1.25 ml chloroform are added to 1 ml of tissue homogenate, followed by vortexing for 2 min. The resulting single-phase extract is chilled for 15 min in an ice bath, and 1.25 ml chloroform and 1.25 ml water are added. After additional vortexing, the phases are allowed to separate. Precipitated proteins accumulate at the interface between the upper (aqueous) and the lower (organic) phase. In many cases phase separation is incomplete. If so, the mixture is centrifuged for 10 min at 5000*g*. The lower phase is recovered, the solvent is evaporated, and the dry lipids are reconstituted in a small volume of methanol/chloroform (1 : 1, by vol.).

24.2.2 Solid Phase Extraction

Lipids can be extracted from large volumes of biological fluids such as serum, urine, cerebrospinal fluid, and so on by solid phase extraction (Figure 24.3). This simple technique for an initial purification of lipids is based on the principle that lipids and other hydrophobic products bind to a hydrophobic matrix. Typically the hydrophobic matrix (C_{18} modified silica gel) is loaded into small cartridges, through which the biological fluid is then pressed. Lipids adhere to the matrix and are then eluted with small volumes of organic solvents. A large variety of chemically modified silica matrices (C_8-, C_{18}-, phenyl-, NH_2-, and aminopropyl-modified silica gels) are available for solid phase extraction of lipids.

Besides lipid extraction, solid phase extraction systems can also be used for crude fractionation of lipid classes. For instance, from a lipid extract of human plasma, the lipid

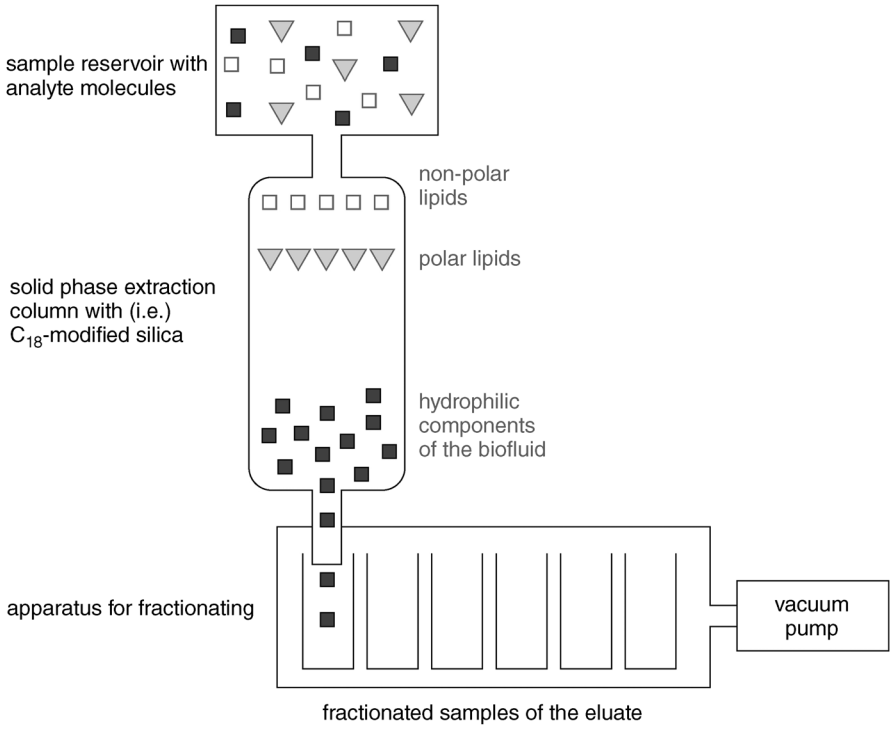


Figure 24.3 Solid phase extraction of lipids from biological materials. Because of their high affinity for reversed phase materials (C_{18}-modified silica), lipids can be extracted from biological fluids. While hydrophilic components of the biofluid do not interact with the matrix material, the more hydrophobic lipids are retained. Using this method, lipids can be concentrated from large volumes of fluid. In addition, various lipid classes can be separated based on their polarity.

Table 24.1 Solvent mixtures for sequential elution of lipids.

Lipid class	Solvent	Solvent ratios (by vol.)
Acidic phospholipids	Hexane/propanol/ethanol/0.1 M ammonium acetate/water/formic acid	420 : 350 : 100 : 50 : 0.5
Neutral phospholipids	Methanol	1
Free fatty acids	Diethyl ether/glacial acetic acid	100 : 2
Neutral lipids	Chloroform/isopropanol	2 : 1

classes listed in Table 24.1 can be isolated by sequential elution with 4 ml of the corresponding solvent mixture from an aminopropyl Bond Elute cartridge. It should be emphasized however, that extraction cartridges supplied by different suppliers have variable extraction properties. It is therefore important to ascertain that a protocol taken from the literature can be reproduced under the given laboratory conditions.

During liquid lipid extraction large volumes of dilute extracts are generated. A rotary evaporator can be used to remove the solvent and concentrate the solution. Foam formation during the evaporation process can be reduced by adding small amounts of ethanol. If greater amounts of glacial acetic acid are present in the mix, it can be removed by azeotropic distillation with toluene.

Lipids should not be stored as dry lipids but as concentrated solutions in glass vials with Teflon-lined caps. Solvent-resistant polypropylene flasks can in some cases also be used. Note, however, that during storage plastic softeners might leach from the flask wall into the lipid extract, causing contamination of the sample. Long-term storage of lipids should be at low temperatures (−80 °C) in the absence of light and oxygen (argon atmosphere). The addition of antioxidants (0.01% 3,5-di-*t*-butyl-4-hydroxytoluene, BHT, relative to amount of lipids) may increase the stability of lipid preparations. Methanol/chloroform mixtures (e.g., 1 : 9 for phospholipids), hexane (for neutral lipids), or pure methanol are suitable storage solvents, but diethyl ether should not be used.

24.3 Methods for Lipid Analysis

24.3.1 Chromatographic Methods

The dominant method for lipid analysis is chromatography, particularly liquid chromatography (LC) and gas chromatography (GC). During the past 30 years the introduction of high performance thin-layer chromatography (HPTLC), the development of high performance liquid chromatography (HPLC), and of capillary gas chromatography have considerably advanced the field of lipid analysis. Through chemical modification of purified silica gel, which constituted the standard matrix for liquid chromatography for a long time, and through the synthesis of novel stationary phases a multitude of new chromatographic matrices has been developed. This virtually unlimited multiplicity of stationary phases strongly advanced the chromatographic separation of critical pairs of analytes. With specifically designed chiral stationary phases, it is even possible to separate enantiomers, without converting them into diastereomers, by the use of chiral resolving agents. In modern liquid chromatography simple modifications of either stationary or/and mobile chromatographic phases makes it possible to solve virtually all separation problems.

Ion Pair Chromatography, Section 10.4.2

Liquid Chromatography For the separation of neutral lipids, adsorption chromatography on silica-based matrices is the method of choice. In contrast, charged lipids are typically analyzed by ion pair chromatography. Here, ionic lipids form complexes with oppositely charged counter-ions, which provide the mobile phases. These internal complexes can then interact with the stationary phase of the chromatographic system. The charge state of the lipid, and with it the ion pair formation, depends on the pH of the mobile phase. Changing the pH thus allows for modification of the chromatographic properties of the analytes. In general, optimization of an elution profile can be achieved by (i) changing the stationary phase or (ii) changing the mobile phase.

In most cases, the strategy for chromatographic analysis of biologically relevant lipids should start with RP-HPLC. This method is easy to use and is less susceptible to disturbance than normal phase HPLC. In fact, NP-HPLC is often subject to technical problems like column clogging, unstable baseline, and solubility issues.

Changing the Stationary Phase In the early stages of development of liquid chromatography silica gel and aluminum oxide were most frequently employed as stationary phases. Later advances included the targeted chemical modification of silica gel, as well as the use of synthetic polymers as sorbents. Today a large variety of stationary phases are commercially available and each of them exhibits unique separation properties. Currently, modified silica gels are the most widely used stationary phases, although synthetic polymers are catching up.

Non-modified silica gel is a polar chromatographic matrix with variable water content. To improve the separation properties of silica gel plates the water should be removed by heating the plates to 150 °C. This dehydration is particularly important for thin-layer chromatography of lipids in non-aqueous solvent systems. Chromatography with non-modified silica is known as normal phase liquid chromatography (NP-LC) and must be distinguished from reversed phase chromatography (RP-LC).

In reversed phase chromatography, the free OH residues of silica are modified by functional groups such as aromatic rings or long alkyl chains, which leads to a reversal of polarity (a nonpolar chromatographic matrix). Both normal phase- and reversed phase-chromatography are used in lipid analysis mostly as a HPLC technique (*reversed phase* (RP)-HPLC and normal phase (NP)-HPLC), and they can be combined to solve specifically challenging separation problems.

Reversed Phase Chromatography, Section 10.4.2

Normal Phase Chromatography, Section 10.4.3

Changing the Mobile Phase For the analysis of most lipid classes commercially available normal phase or reversed phase columns are sufficient. The optimization procedure in the laboratory therefore mostly consists of modifying the mobile phase. Depending on the type of solvent system, isocratic separations (using one solvent system with specific composition) are distinguished from gradient separations (where the solvent system is modified over time). For complex column chromatographic separations binary, ternary, and quaternary gradient elution systems are employed.

Thin-Layer Chromatography Nowadays, thin-layer chromatography (TLC) is typically carried out using precast silica gel plates. On these plates the stationary phase is provided as a

thin layer of silica gel on top of a flat sheet of aluminum or plastic foil or glass plates. The lipids to be analyzed are usually dissolved in a volatile solvent such as hexane or chloroform. Water should be removed from the lipid mixture by azeotropic distillation, because it complicates analysis by modifying the binding properties of the stationary phase. The lipid mixture can then be applied as a thin band in a labeled area of the silica plate using a micro-liter syringe, or a capillary tube. Commercial applicators are available to facilitate the application process. There should empty space of at least 0.5–1 cm between the samples to prevent sample mixing during the run. After the application the solvent is completely evaporated, the plate is placed in a tank that is partially filled with solvent. It is important to have a sufficient solvent level to completely immerse the lower edge of the plate. On the other hand, the solvent should not reach the applied samples as this would lead to the samples being eluted into the solvent reservoir. Usually, the chromatograms are developed under saturating conditions, which means that the atmosphere in the chromatographic chamber is saturated with solvent. Homogeneous saturation of the atmosphere in the chamber can be achieved by inserting a solvent saturated sheet of filter paper in the chromatographic chamber. For special separation problems non-saturating atmospheres may be preferred.

After the thin-layer plate is placed in the chromatographic chamber, the solvent begins to move upwards in the stationary phase, driven by capillary forces. The various lipids are transported along with the solvent depending on their affinity for the stationary phase. For example, when a mixture of neutral lipids is separated on a polar silica gel plate, the more polar monoacyl glycerols are retained more strongly than the nonpolar triacyl glycerols.

Thin-layer chromatography is usually terminated when the solvent front has reached a virtual line, which is about 1–2 cm from the top end of the thin-layer plate. In slow running solvent systems (usually those containing water) chromatography should be stopped earlier, because at long running times the chromatographic bands tend to broaden considerably.

There are several methods for the visualization and quantization of thin-layer chromatograms; for lipids, fluorescence densitometry is the most useful. The developed chromatogram is impregnated with a solution of a fluorescence indicator, such as 1 mM 6-*p*-toluidino-2-naphthalene-sulfonic acid (dissolved in 50 mM Tris-HCl). Following illumination with a UV lamp, the dye emits light in the visible range if it is located in a lipophilic environment. Thus, lipid bands light up. Because the fluorescence decays quickly, a photograph should be taken immediately following exposure. The photograph can then be evaluated by quantitative densitometry. Another way of visualizing lipid bands is by staining with 4-(*N*,*N*-dihexadecyl) amino-7-nitrobenz-2-oxa-1,3-diazole (NBD-dihexadecylamine). This lipid-soluble fluorescent dye is added to the solvent at concentrations of 0.02–0.05%; the solvent distributes the dye over the whole plate during the chromatographic run. The resulting fluorochromatogram can then be quantified directly by reflection or transmission densitometry. Many commercial TLC plates already contain fluorescent indicators as a component of the stationary phase. Other staining methods include reversible iodine staining or carbonization with sulfuric acid at 200 °C.

TLC allows for the determination of R_f values for different components of a complex mixture. This value is defined as the ratio of the migration distance of the substance divided by the migration distance for the solvent front. The R_f value corresponds to the retention volume or the retention time in column chromatography. Notably, the R_f values are not invariable constants for different substances. In contrast, R_f values strongly depend on the chromatographic conditions. The R_f values for different substances are therefore only strictly comparable if they have been determined under identical chromatographic conditions (on the same TLC plate). Besides the classical single development method, special development techniques have been worked out to optimize the separation performance of TLC.

Multiple Development In conventional multiple development, the chromatogram is first developed with a solvent system with high elution power. The solvent is then evaporated, and a second chromatography is performed in the same direction using a solvent system with low elution power. This procedure may be repeated several times. The advantage of multiple development is that for critical pairs of analytes improved resolution is achieved, because the substance bands are concentrated in each additional chromatography step and thus band broadening is counteracted even when relatively high R_f values are reached.

In the automated version of multiple development (AMD, automated multiple development) the chromatogram is divided into 20–30 segments. Each of the consecutive segments is 2–4 mm longer than the previous one. Following sample application, the chromatogram is developed in a solvent system with high elution power until the solvent has reached the top of the first segment. After the plate has dried, it is then developed with a solvent system of slightly lower elution power. Once the solvent has reached the top of the second segment, it is dried and then

developed again with an elution system of lower elution power than the second one. The major advantage of AMD is that it narrows the bandwidths (0.2–2 mm) of the different substances, which concentrates the substance in a narrow band. This improves the detection limits. Furthermore, the resolution is dramatically enhanced compared to single step development. Through complete automation this procedure has become much less work intensive.

Two-Dimensional Thin-Layer Chromatography In two-dimensional thin-layer chromatography, the first chromatography step serves to achieve a crude separation of lipid classes (e.g., separation of phospholipids and neutral lipids). Following solvent evaporation, the plate is turned 90° and developed with a second solvent system, which is suitable for fractionation of each specific lipid class (e.g., subdivision of phospholipids). The drawback of two-dimensional thin-layer chromatography is that a whole chromatography plate is needed for each lipid extract. Thus, direct comparison of two samples on one plate is hardly possible.

Chromatographic Separation Techniques, Chapter 10

Column Chromatography In contrast to thin-layer chromatography, the stationary phase in column chromatography is packed into a column and the mobile phase continuously flows through the particles of the stationary phase. Pre-packed HPLC columns are commercially available from several vendors, so that self-packed columns are rarely used nowadays. In addition to the appropriate columns, which are packed with stationary phase suitable for the specific separation problems, HPLC requires pumping systems to provide constant solvent flow, a gradient mixer, and suitable detectors, which are coupled serially or in parallel.

Common stationary phases for lipid analysis include non-modified silica gels (normal phase HPLC, NP-HPLC) as well as silica gels modified with the following functional groups: cyanopropyl, aminopropyl, diol, silver ion (for normal phase separations); C_{18}, C_8, phenyl (for reversed phase separations); DEAE or TEAE (for ion exchange columns); cyclodextrin or Pirkle phases (for separation of enantiomers). A new development is the use of *narrow bore* columns in HPLC. Because of the fine granules of the packing material (<3 µm) and the small column diameter (1 mm), miniaturized columns can be used without sacrificing the resolution of separation. This significantly reduces the solvent volume (flow rates 20–200 ml min^{-1}) and even improves the separation power of specific applications.

In normal phase chromatography (polar silica gel as stationary phase) nonpolar solvent systems (e.g., *n*-hexane/2-propanol or chloroform/methanol mixtures) are predominantly used as mobile phase. In this case, substances are eluted in order of increasing polarity (nonpolar lipids first). In reversed phase settings with nonpolar C_8- or C_{18}-phases, solvent mixtures with methanol, acetonitrile, and water are commonly employed. Here substances elute in order of increasing hydrophobicity (hydrophilic lipids first). In ion pair chromatography, which in lipid analysis is used for the separation of charged phospholipids, counter-ions (e.g., triethylamine, choline chloride, or buffer ions) are added to the solvent mix.

A special problem of lipid analytics is the detection of lipids following separation. Commercial HPLC systems frequently involve UV detectors, which are of only limited use in lipid analysis. Only lipids containing aromatic residues or conjugated double bonds are readily detected due to their characteristic UV-chromophores. Lipids with unsaturated fatty acids can still be detected at 205 nm. Some lipid classes that do not contain characteristic chromophores can be derivatized (either pre- or post-column) with chromophore or fluorescent groups if reactive coupling groups such as free carboxylic or hydroxyl groups are available in the lipids.

As an alternative to UV detection, refraction index detectors can be used in lipid analysis. Inclusion of evaporative light scattering detectors (ELSD) in lipid analysis may improve lipid analysis in the future. In these detectors, the solvent is quickly evaporated after separation in the detector chamber. The non-volatile lipids remain as condensed droplets in the detector and these droplets diffract incoming laser light. Because the amount of condensed lipid is proportional to the level of diffraction, chromatograms can be accurately quantified using this method.

In recent years, flame ionization detectors (FID) have been introduced to HPLC. As in ELSD detectors, the solvent needs to be removed prior to the ionization. When working with radioactive lipids fractionation of the column effluent and subsequent discontinuous liquid scintillation counting or continuous online radioactivity measurement can be employed. Continuous online radioactivity detectors are equipped either with solid scintillators or with a radioactivity flow-through chamber, in which a liquid scintillator is mixed with the column

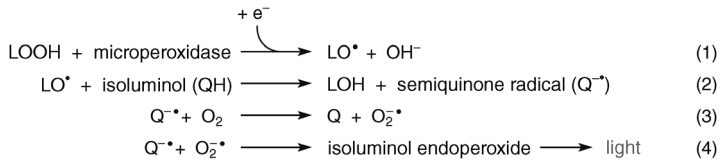


Figure 24.4 The principle of chemiluminescence detection of hydroperoxy lipids. Following an admixture of isoluminol and microperoxidase to the column effluent, the indicator reaction proceeds as shown. The emitted light can be quantified as a measure of peroxide concentration. LOOH = hydroperoxy lipid, $LO^{\bullet}$ = alkoxy radical, and $O_2^{\bullet -}$ = superoxide radical.

effluent. Solid scintillators are easier to use, but are less sensitive than liquid scintillators. Online liquid scintillation counting might be problematic in some cases since mixing the column effluent with the scintillation fluid might precipitate the scintillator.

Under an oxygen atmosphere unsaturated fatty acids are oxidized to hydroperoxy derivatives and in some applications quantifying these hydroperoxy lipids is of biological interest. Addition of a solution of isoluminol plus microperoxidase to the column effluent induces chemiluminescence in the chromatographic fractions containing hydroperoxy lipids. The changes in chemiluminescence can be measured with an online luminometer connected to the column outlet (Figure 24.4).

Gas Chromatography In contrast to liquid chromatography, gas chromatography employs a carrier gas as mobile phase. Only lipids that vaporize at up to 350 °C without degradation can be analyzed by gas chromatography. Among these are derivatized fatty acids, sterol derivatives, derivatized mono- and diacyl glycerols, as well as triacyl glycerols. The carboxy group of free fatty acids needs to be methylated prior to analysis. The OH groups are silylated, if necessary, by treatment with bis(trimethylsilyl)trifluoroacetamide (BSTFA) or a similar silylating agent. Ketones and aldehydes are converted into the corresponding methoximes by reacting with methoxyamine hydrochloride ($CH_3ONH_2 \cdot HCl$) in pyridine. Since interaction of the analytes with the stationary phase is temperature dependent, the column temperature can be varied to modify the retention characteristics. Therefore, the column temperature in gas chromatography is usually varied with a linearly or nonlinearly increasing gradient. This temperature gradient serves a similar role in gas chromatography as the solvent composition gradient in liquid chromatography.

For gas chromatographic lipid analysis, both packed columns and capillary columns can be used. Packed columns are typically 1–2 m long, with an inner diameter of 2–4 mm. They are filled with an inert matrix coated with the separation phase. For the separation of lipids, mainly nonpolar thermostable silicon elastomers such as OV-1, OV-101, SP2100 or SE-30 at a concentration of 1–3% are employed. Self-packing of well-functioning GC-columns requires great experience and frequently fails.

Capillary columns usually have superior separation performance when compared with packed columns. They are around 15–30 m long, have an inner diameter of 0.15–0.5 mm, and are coated on the inside with a thin film (0.1–5 mm) of a stationary phase. Columns with thin film thickness are particularly useful for the analysis of compounds with high boiling points, such as fatty acids and triacyl glycerols. For lipid analysis, typically, separation phases consisting of polysiloxane derivatives are used. They exhibit low polarity and high thermostability. Because most lipids have very high boiling points and need to be separated at high temperatures, the thermostability of the separation phase is a major point to be considered. If the characteristics of the lipids to be analyzed cover a wide polarity range it is useful to initially employ weakly polar columns (universal columns) such as 5% phenyl siloxane. Most manufacturers offering capillary columns provide application recommendations, as to which kind of column may be most suitable for a given separation problem.

Nonpolar gas chromatographic columns separate lipids mainly according to the number of carbon atoms they carry. In contrast, for polar and semi-polar columns additional structural parameters (e.g., number, position, and geometry of double bonds) more strongly impact the separation profile. Since the retention times in gas chromatography depend on the nature of the column, the C-value was introduced to characterize unknown lipid components. In this approach, the chromatographic column is standardized for each series of analyses by running a reference mix of saturated hydrocarbons as standards (Figure 24.5). Hexane elutes with a *C*-value of 6, dodecane with a *C*-value of 12. For linoleic acid we determined a *C*-value of 21.1.

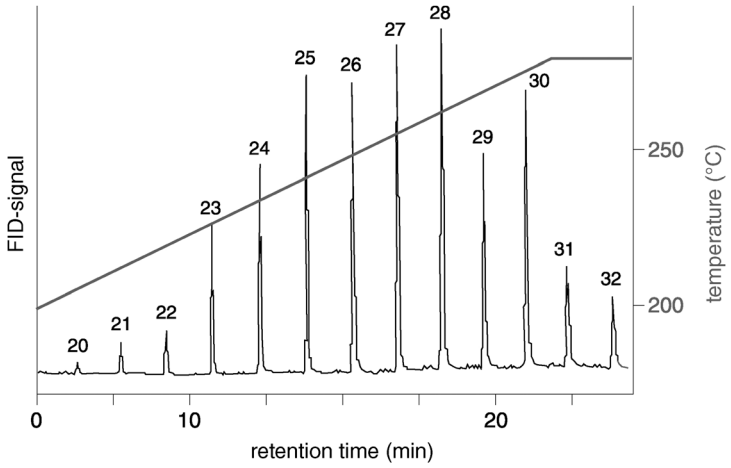


Figure 24.5 Standardization of a gas chromatographic system with a mixture of long-chain hydrocarbons. To determine the *C*-values of specific analytes, a mixture of long-chain hydrocarbons (C_{20} to C_{34}) is separated on a HP-1 capillary column (12 m × 0.2 mm, film thickness 0.33 mm). Temperature profile: 2 min at 180 °C, followed by a gradient of 5 °C min^{-1} up to 280 °C, which is then maintained for 15 min. The numbers above the peaks denote the number of C atoms of the saturated hydrocarbons.

24.3.2 Mass Spectrometry

Mass Spectrometry, Chapter 15

Similar to gas chromatography, for mass spectrometric analysis the lipids to be analyzed need to be evaporated. In most cases, this is achieved by raising the temperature and by lowering pressure. Lipids with high boiling points, such as charged phospholipids, can be detached from a solid matrix either by bombardment with high-energy particles (fast atom bombardment, FAB) or by using a laser beam (MALDI).

To analyze non-charged lipids by mass spectrometry they need to be ionized following evaporation. Currently used ionization procedures in lipid analytics are electron impact ionization (EI), chemical ionization (CI), electrospray ionization (ESI), and thermospray ionization (TSI). In EI, the lipid molecules are bombarded under vacuum with high-energy (around 70 eV) electrons, which fragment the lipids into a characteristic pattern of ions. In CI, a rather soft ionization method, little fragmentation occurs so that mainly charged molecule-ions are generated. A special form of CI that is used in lipid analysis *inter alia* to quantify eicosanoids, is called electron capture mass spectrometry. When high-energy electrons interact with a gas in the ionization chamber they are sufficiently slowed down to be captured by electrophilic functional groups of the analyte without causing fragmentation. This transforms non-charged analyte molecules into negatively charged molecule ions. Since naturally occurring lipids rarely involve such highly electrophilic functional groups, they must be introduced by derivatization of the analytes. For example, reacting the free carboxylic groups of fatty acid derivatives with pentafluorobenzoyl chloride generates the corresponding fatty acid pentafluorobenzoyl esters, which can easily capture slow electrons. Since fragmentation rarely occurs, this technique is useful to determine the molecular weights of the analytes. In contrast to EI and CI, which require the reaction to proceed under vacuum, ESI is carried out at atmospheric pressure (Section 15.2) and can therefore be readily combined with liquid chromatography (Section 24.3.5). Electrospray ionization mass spectrometry (ESI-MS) is the method of choice for the characterization of cellular lipidomes (Section 24.6).

24.3.3 Immunoassays

Immunoassays, Section 5.3.3

To quantify lipid hormones (e.g., steroids, eicosanoids) in biological fluids, immunoassays are commercially available that in principle feature high sensitivity and specificity. The high specificity is achieved through the use of monoclonal antibodies, the great sensitivity is due to the highly sensitive detection system. Different types of immunoassays are available and these include radioimmunoassays, enzyme immunoassays, fluorescence immunoassays, and chemiluminescence assays. Several companies offer complete kits to quantify various lipid hormones, kits that contain all the necessary reagents including quantification standards. A critical point, however, is that the specificity of the monoclonal antibodies employed has often not been characterized sufficiently and cross-reactivity with other antigens can never been completely ruled out. Immunological quantification of certain eicosanoids has frequently led to

significantly higher values than their quantification with physicochemical methods (HPLC, LC-MS/MS).

24.3.4 Further Methods in Lipid Analysis

Lipids predominantly contain carbon, hydrogen, oxygen, nitrogen, and phosphorus; other elements are rarely present. Quantitative elemental analysis can therefore usually be restricted to these elements. The data from elemental analysis allow important conclusions to be made about the structures present in purified lipid fractions. The purity of the lipid fraction is crucial to the success of elemental analysis. Highly pure lipids can be prepared by repeated rounds of HPLC purification using different stationary phases, such as consecutive RP- and NP-HPLC. Of special importance for the quantification of phospholipids is the determination of phosphorus in organic lipid extracts (Section 24.4).

Occasionally infrared spectroscopy (IR) is used for structure determination of purified lipid fractions. Although the information content of the spectra is limited, they do present evidence for the presence of specific functional groups (e.g., free hydroxy or keto groups). Conventional IR spectrometers do not have sufficient sensitivity for bioanalysis. However, using Fourier transform IR spectroscopy (FT-IR), interpretable spectra in the µg range can be achieved. A classical method for the determination of the percentage of *trans* fatty acids in lipid extracts is based on the quantification of the IR band between 900 and 1000 cm^{-1}.

IR Spectroscopy, Section 7.3

UV/Vis spectroscopy plays a relatively minor role in general lipid analysis, but is of great importance for the analysis of lipid peroxidation products. The biologically important polyunsaturated fatty acids (PUFAs) contain two more 1,4-*cis,cis*-dienes, which in the course of lipid peroxidation are converted into cis, trans-, or trans,trans-conjugated systems. These conjugated dienes have a characteristic UV spectrum with an absorption maximum between 230 and 235 nm and a relatively high molar extinction coefficient (25 000 M^{-1} cm^{-1}). Besides conjugated dienes, the oxidation of higher unsaturated fatty acids (octadecatrienoic acid, eicosatetraenoic acid) also leads to conjugated trienes, conjugated tetraenes, and conjugated ketodienes, which also carry characteristic UV-chromophores (Table 24.2).

UV/Vis Spectroscopy, Section 7.2

^{1}H, ^{13}C, and ^{31}P NMR spectroscopy have gained increasing importance in the structural analysis of lipids. These techniques provide detailed information about the position and the stereochemistry of C=C double bonds in fatty acids, as well as about the nature and position of functional groups (Figure 24.6). The increasing role of NMR spectroscopy in lipid analysis is due to the development of powerful NMR spectrometers with strongly improved sensitivity, but also to improvements of the chromatographic purification techniques. Today, 600 MHz spectrometers deliver informative lipid spectra with 10–50 µg of lipid. To obtain good NMR spectra, the lipids to be analyzed should be at least 90% pure. This level of purity should be documented in several chromatographic systems (e.g., RP-HPLC, NP-HPLC, and GC). A special challenge for NMR analysis of lipids is the fact that many different isomers may exist (positional isomers, stereoisomers, etc.) that are sometimes difficult to resolve by chromatography, yet deliver distinct NMR signals. High isomeric purity is thus a requirement for obtaining informative NMR spectra. This can be achieved by a combination of consecutive chromatographic steps including chiral purification.

NMR, Chapter 18

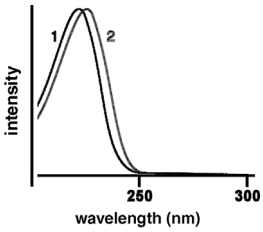
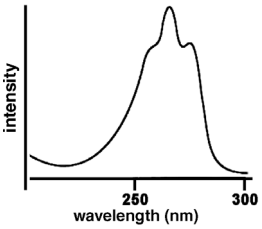
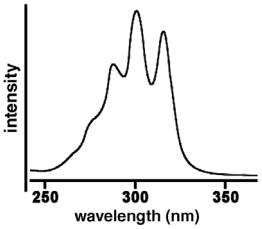
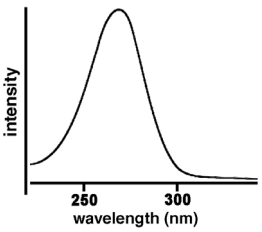
24.3.5 Combining Different Analytical Systems

Great progress in lipid analysis has been made in the last decade by sequentially combining different analytical techniques. An important prerequisite for this development was the extensive progress made in data processing and computer technology, without which the massive amounts of data generated by these combinations could hardly be handled.

Coupling of HPLC and UV/Vis Spectroscopy The development of diode array detectors has allowed the recording of complete UV/Vis spectra at any given time point of a chromatogram without interrupting the chromatographic process. The data obtained by diode array detectors during a chromatographic run can be visualized as a three-dimensional chromatogram (Figure 24.7a), with the retention time on the *x*-axis, light absorption on the *y*-axis, and

Chromatographic Separation Techniques, Chapter 10

Table 24.2 UV chromophores of oxidized fatty acids.

Chromophore[a)]	λ_{max} (nm)	ε (M^{-1} cm^{-1})	Spectrum
Conjugated diene (–C=C–C=C–) e.g., 15-HETE	231 (*F,E*) (curve 1) 235 (*Z,E*) (curve 2)	24 000	
Conjugated triene (–C=C–C=C–C=C–) e.g., LTB_4	268	32 000	
Conjugated tetraene (–C=C–C=C–C=C–C=C–) e.g., LXB_4	301	50 000	
Conjugated ketodiene (–C(=O)–C=C–C=C–) e.g., 13-KODE	270	24 000	

a) HETE: hydroxyeicosatetraenoic acid, LT: leukotriene, LX: lipoxin, KODE: 13-keto-*cis,trans*-$\Delta^{9,11}$octadecadienoic acid.

wavelength on the *z*-axis. In lipid analysis the diode array detector is used especially for the determination of oxidized lipid species, because these oxidized derivatives contain conjugated double bonds (dienes, trienes, tetraenes), giving rise to characteristic absorption spectra.

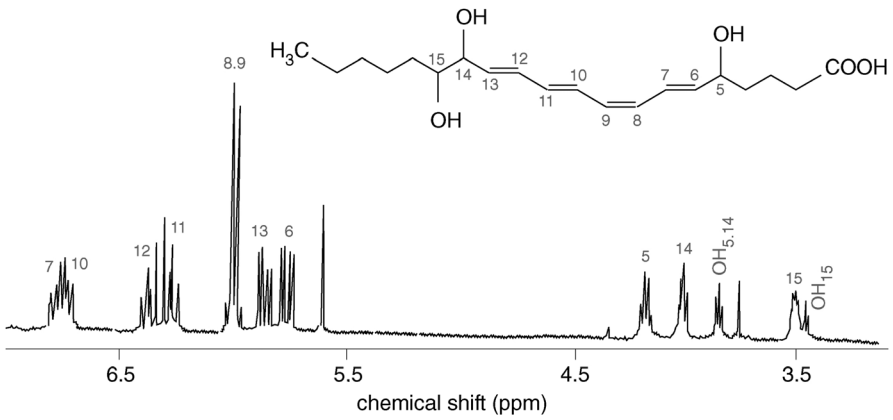


Figure 24.6 Segment of the ^{1}H NMR spectrum of (5*S*,14*R*,15*S*)-trihydroxy–(6*E*,8*Z*,10*E*,12*E*)-eicosatetraenoic acid (lipoxin B_4). The (*cis, trans*) geometry of the double bonds was concluded from the coupling constants of the olefinic protons.

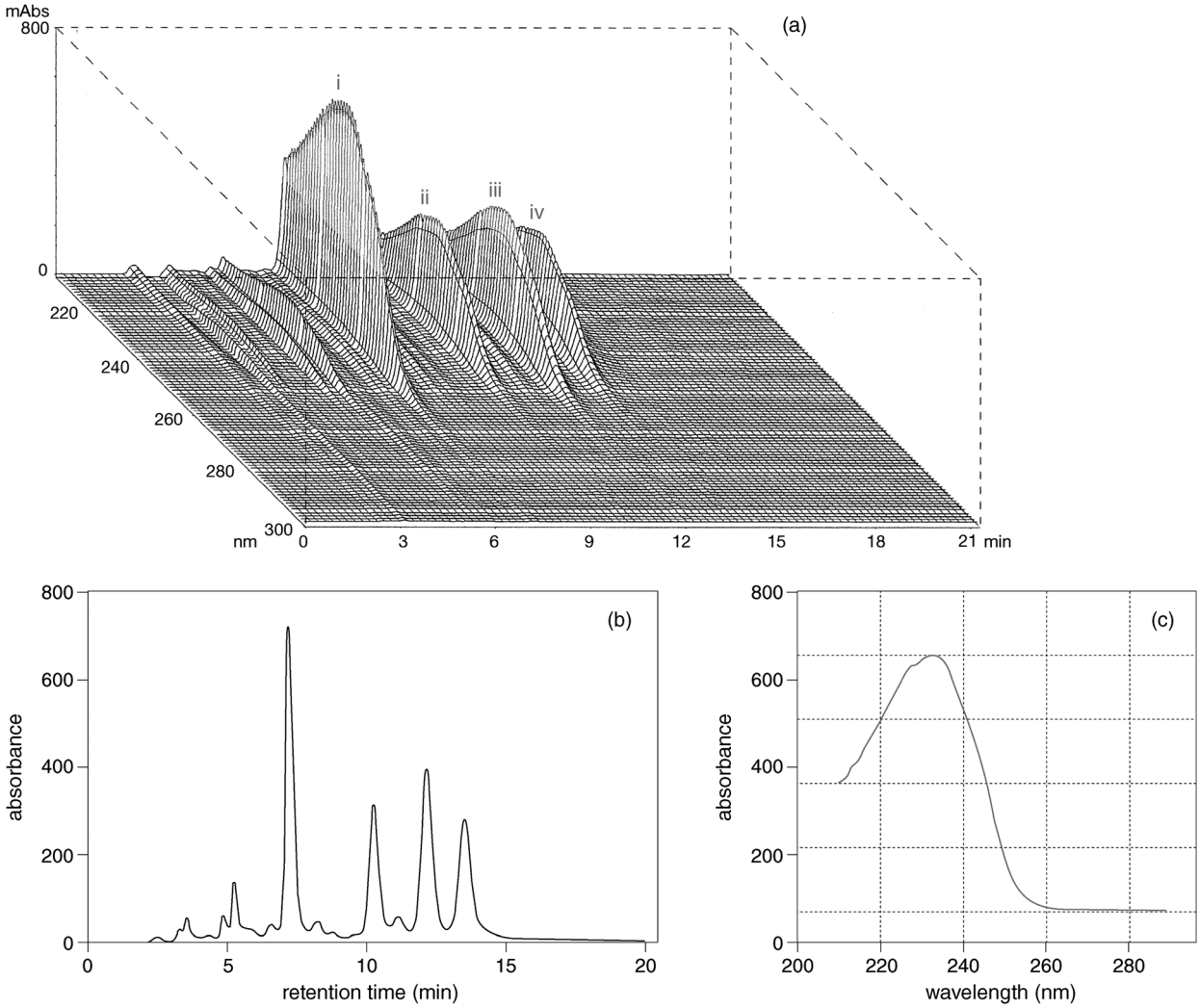


Figure 24.7 Three-dimensional chromatogram of a mixture of hydroxy linoleic acid isomers obtained using a diode array detector. The mixture of hydroxy linoleic acids was separated by NP-HPLC with a solvent mixture of *n*-hexane/2-isopropanol/glacial acetic acid (100 : 2 : 0.1, by volume) on a Nucleosil-50 column (250 mm × 5 mm, 7 µm particle size). (a) Three-dimensional chromatogram; (b) two-dimensional chromatogram (section from three-dimensional chromatogram taken parallel to the *x*-axis at a wavelength of 235 nm); (c) UV-spectrum of the main component (section from the three-dimensional chromatogram taken parallel to *y*-axis at 7.3 min). (i) 13-Hydroxy-*cis,trans*-$\Delta^{9,11}$-octadecadienoic acid, (ii) 13-hydroxy-*trans, trans*-$\Delta^{9,11}$-octadecadienoic acid, (iii) 9-Hydroxy-*trans,cis*-$\Delta^{10,12}$-octadecadienoic acid, and (iv) 9-Hydroxy-*trans,trans*-$\Delta^{10,12}$-octadecadienoic acid.

Coupling of HPLC and Mass Spectrometry (LC/MS) With the introduction of electrospray ionization, the combination of LC with MS became feasible. This method can in principle detect all charged or non-charged lipids, as long as they can be ionized by the electrospray procedure. To prevent damage to the LC-MS interface and to protect the ion source of the MS, which frequently occur at typical HPLC flow rates of 1 ml min^{-1}, the solvent stream is split following column elution, with only 10% (100 µl min^{-1}) delivered into the ionization chamber. As an alternative to splitting the eluent, narrow bore columns can be used for HPLC, which are run at lower flow rates and are suitable for most lipid separations.

Coupling of Gas Chromatography and Mass Spectrometry (GC/MS) Because capillary gas chromatography uses very low flow rates for the carrier gas, the effluent of a capillary column can be directly introduced into the ion source of a mass spectrometer via an interface without disrupting the vacuum present in the mass spectrometer. Thus, all lipids analyzed by gas chromatography can also be explored by mass spectrometry for more detailed structural information. Since most lipids are low molecular weight compounds sophisticated mass spectrometers are not required for lipid analysis. Simple mass sensitive detectors are sufficient in many cases. Such

Mass Spectrometry, Chapter 15

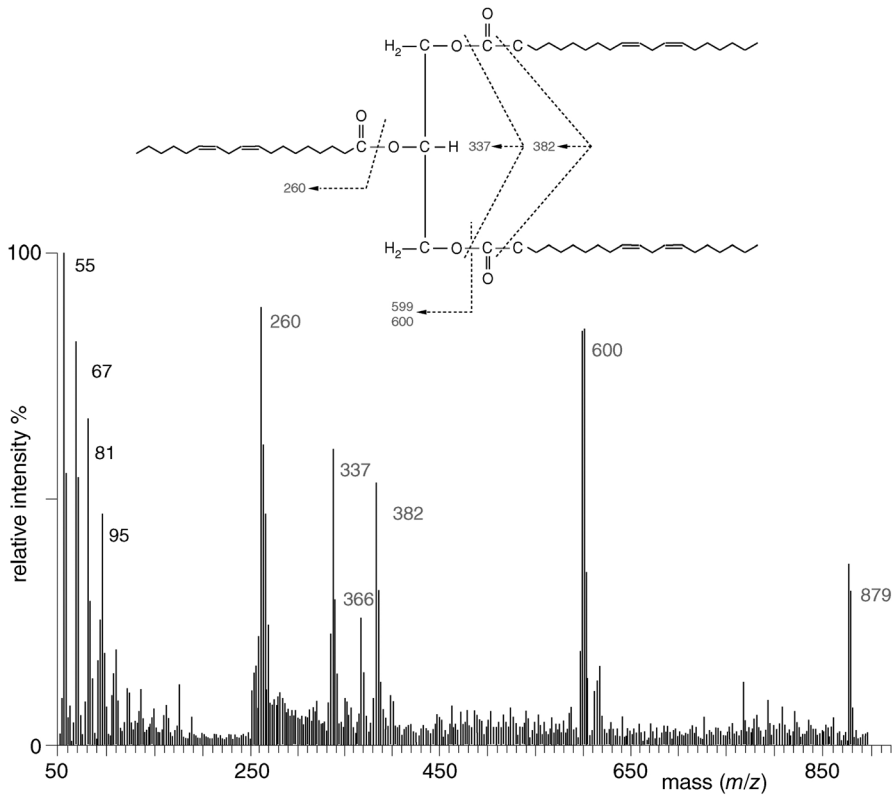


Figure 24.8 Mass spectrum of trilinolein. Information about the structure of the detected ions can be derived from the fragmentation pattern. The ionized molecule exhibits an *m/z* ratio of 879, but its protonated form at *m/z* 880 is also detected. The base peak in the informative range of the spectrum represents the RO^+ ion. EI of unsaturated triacyl glycerols typically leads to a loss of three hydrogen atoms from the RO^+ ion (theoretically *m/z* 263, observed *m/z* 260).

mass sensitive detectors are quadruple mass spectrometers that work, usually, with electron impact ionization and are limited to a mass range of up to 1000 Da. Despite these limitations, mass sensitive detectors are suitable for detailed structural elucidation of simple lipid species (Figure 24.8).

MS/MS Section 15.2

Tandem Mass Spectrometry (MS/MS) In tandem mass spectrometry, two mass spectrometers are coupled, with the first mass spectrometer providing the selection of specific molecular ions (separation according to the molecular mass), which are then fragmented in the second mass spectrometer to deduce structural information from the typical fragmentation pattern. MS/MS coupling can readily be used to analyze charged or easily chargeable lipids (e.g., phospholipids). For analysis of neutral lipids (e.g., triacyl glycerols, cholesteryl esters) the appropriate ionization conditions must first be determined. For example, triacyl glycerols form anionic complexes with lithium ions, and can thus easily be analyzed by ESI-MS. MS/MS is of special importance for the analysis of complex lipid mixtures (cellular lipidomes); it can be performed on Folch extracts of plant or animal tissues without prior chromatographic separation but also in connection with LC.

24.4 Analysis of Selected Lipid Classes

24.4.1 Whole Lipid Extracts

Lipid extracts of biological samples contain a large number of structurally distinct lipid classes, which differ from each other with respect to their polarity. For instance, retinols and cholesterol esters are very nonpolar lipids, which intensively interact in reverse phase chromatography with the chromatographic matrix. In contrast, lysophosphatides and complex glycolipids (gangliosides) are rather hydrophilic and, thus, only weakly interact with reverse chromatographic matrices. Thus, it is a rather difficult task to develop chromatographic systems for the complete separation of complex lipid extracts into their major components. For this purpose complex gradient systems are required, which cover a large polarity range. In critical segments of such chromatograms isocratic regions must be included in order to separate analytes with similar

polarities. For instance, for the separation of complex lipid extracts on silica gel based chromatographic matrices tertiary or even quaternary gradients consisting of isooctane, *n*-hexane, 2-propanol, tetrahydrofuran, and water are frequently employed. Since native lipids usually do not carry strong chromophores the use of UV/Vis-detectors is restricted to specific applications (oxidized lipids). In most other cases mass-sensitive or electrochemical detectors (Section 24.3.2) may be used. Employing electrospray ionization and tandem mass spectrometry (ESI-MS/MS) the composition of complex lipid extracts can be analyzed in a single analytical run (Section 24.6).

24.4.2 Fatty Acids

Fatty acids are important structural components of many lipids (Table 24.3) and, thus, analysis of the fatty acid composition of a lipid extract constitutes a key element in the analytical hierarchy of complex lipids. Fatty acids in highly developed organisms are monocarboxylic acids with mostly even numbers of carbon atoms (4–22 carbon atoms) and 0–6 *cis* double bonds. *Trans* fatty acids are rare but exhibit unfavorable biological properties (proatherogenic) if present in nutrient lipids. The method of choice for fatty acid analysis is capillary gas chromatography on apolar or medium-polar stationary phases. Since free fatty acids carry a carboxylic group they vaporize only at high temperatures. Thus, gas chromatography can hardly be carried out with free fatty acids. However, after methylation of the carboxylic group the evaporation temperature is greatly reduced, which enables quantitative gas chromatography. Since fatty acids occur in biological samples predominantly as fatty acid esters non-enzymatic alkaline hydrolysis of the ester lipids is usually carried out. For sensitive lipids, which might decompose under such harsh experimental conditions, enzymatic hydrolysis using an excess of lipid hydrolyzing enzymes (e.g., phospholipases) may be performed. The resulting free fatty acids may subsequently be methylated by standard methylation

Table 24.3 Biologically relevant fatty acids.

Number of carbon atoms	Common name	Systematic name	Melting point (°C)
Saturated fatty acids			
4	Butyric acid	*n*-Butyric acid	–5
6	Caproic acid	*n*-Hexanoic acid	–3
8	Caprylic acid	*n*-Octanoic acid	17
10	Capric acid	*n*-Decanoic acid	32
12	Lauric acid	*n*-Dodecanoic acid	44
14	Myristic acid	*n*-Tetradecanoic acid	54
16	Palmitic acid	*n*-Hexadecanoic acid	63
18	Stearic acid	*n*-Octadecanoic acid	70
20	Arachidic acid	*n*-Eicosanoic acid	76
22	Behenic acid	*n*-Docosanoic acid	80
24	Lignoceric acid	*n*-Tetracosanoic acid	84
Monounsaturated fatty acids[a]			
16	Palmitoleic acid	Δ^{9}-Hexadecenoic acid	0
18	Oleic acid	Δ^{9}-Octadecenoic acid	13
18	Vaccenic acid	Δ^{11}-Octadecenonic acid (*trans*)	40
24	Nervonic acid	Δ^{15}-Tetracosenoic acid	42
Polyunsaturated fatty acids[a]			
18	Linoleic acid	$\Delta^{9,12}$-Octadecadienoic acid	–5
18	α-Linolenic acid	$\Delta^{9,12,15}$-Octadecatrienoic acid	–14
20	Arachidonic acid	$\Delta^{5,8,11,14}$-Eicosatetraenoic acid	–49
20	Eicosapentaenoic acid (EPA)	$\Delta^{5,8,11,14,17}$-Eicosapentaenoic acid	–54
22	Docosahexaenoic acid (DHA)	$\Delta^{4,7,10,13,16,19}$-Docosahexaenoic acid	–44

a) The positions of the double bonds are indicated by the superscripts following the Δ sign.

procedures (e.g., addition of etheric diazomethane). Alternatively, fatty acid methyl esters can be formed directly from extracted ester lipids by transesterification with sodium methoxide. To obtain fatty acid methyl esters from sphingolipids acidic methanolysis can be carried out.

The fatty acid composition of complex lipid extracts can also be analyzed by HPLC without methylation of the alkaline hydrolysates. Since naturally occurring free fatty acids do not contain UV/Vis chromophores mass sensitive detectors (Section 24.3) must be employed. Alternatively, the carboxylic group may be derivatized with chromophores or fluorophores. Unsaturated fatty acids absorb light in the near UV-region and thus are detectable recording the absorbance at 205 nm. Unfortunately, the molar extinction coefficient is rather low and, thus, this method is rather insensitive. However, if analysis is aimed at quantifying only the major polyenoic fatty acids in animal tissue (linoleic acid, linolenic acid, arachidonic acid, eicosapentaenoic acid, docosahexaenoic acid) RP-HPLC provides important information.

24.4.3 Nonpolar Neutral Lipids

Mono-, Di-, Triacyl Glycerols and Waxes In the lipid extracts of animal tissues triacyl glycerols and their partial hydrolysis products, the mono- or diacyl glycerols, represent major components of nonpolar ester lipids (Figure 24.9). These lipid classes can easily be separated from each other by thin-layer chromatography. As mobile phases mixtures of *n*-hexane/diethyl ether/acetic acid (70 : 30 : 1, by vol.) or benzene/acetic acid (88 : 12, by vol.) are frequently employed. To separate 1- and 2-monoacyl glycerols boric acid treated silica gel plates may be employed. Since boric acid interacts with vicinal hydroxy groups 2-monoacylglycerols migrate in front of 1-monoacylglycerols. For gas chromatographic separation of triacyl glycerols capillary columns with apolar or medium polar matrices are employed (Section 24.3.1). On apolar matrices triacyl glycerols are separated from each other according to the chain lengths of their fatty acid molecules. On medium-polar matrices other structural parameters such as the number and geometry of the double bonds impact the retention behavior. Because of the high evaporation temperatures of triacyl glycerols and the intense interactions of these compounds with the stationary phases temperatures as high as 350 °C must be employed. Unfortunately, such high temperatures reduce the life span of the columns. Triacyl glycerols containing unsaturated fatty acid, which luckily do not occur frequently in mammalian tissues, are rather labile under these conditions and undergo thermal decomposition. Diacyl glycerols are analyzed under similar conditions but silylation of the free hydroxy group (using BSTFA) prior to analysis is recommended. Monoacyl glycerols should always be silylated before gas chromatographic analysis.

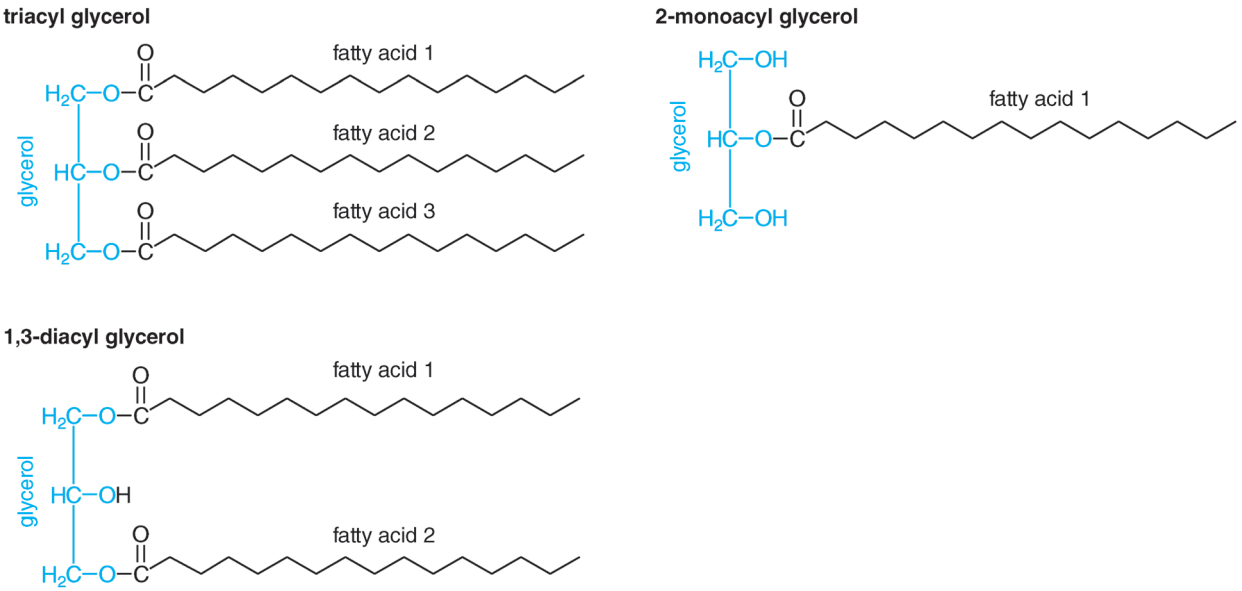


Figure 24.9 Chemical structure of mono-, di-, and triacyl glycerols. These lipids consist of the trivalent alcohol glycerol and three saturated fatty acids. Mono-, di-, and triacyl glycerols carrying unsaturated fatty acids do not occur frequently in animal tissues.

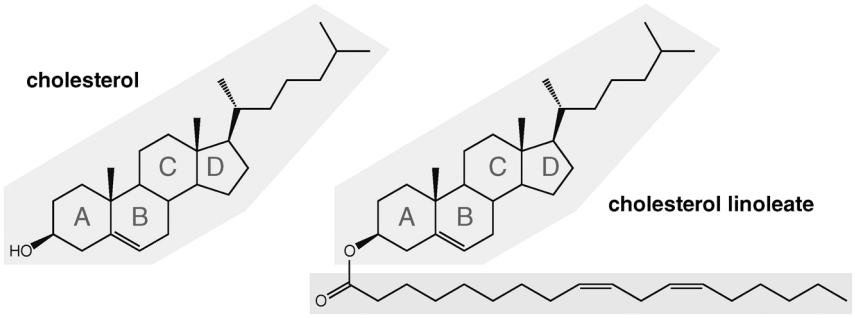


Figure 24.10 Chemical structure of free cholesterol and cholesterol esters. In cholesterol esters, which occur frequently in lipoproteins with low (LDL) and high density (HDL), the free hydroxy group at C3 of the A ring is esterified with a fatty acid such as oleic acid (monounsaturated) or linoleic (polyunsaturated) acid.

For HPLC analysis of mono-, di-, and triglycerides silica gel columns are frequently employed. Mass sensitive detectors (Section 24.3.2) must be used when lipids lacking unsaturated fatty acid are analyzed. If polyenoic fatty acids are present UV-detection at 205 nm is possible but here again the sensitivity of this UV-detection is rather low. The UV-detectability of mono- and diacyl glycerols is strongly improved when chromophores or fluorophores are introduced at the free hydroxy groups.

Waxes are usually analyzed by a combination of sequential chromatographic steps. A frequently employed method for this purpose is ethanolysis with subsequent acetylation of the resulting alcohol component. As a result of this derivatization method waxes are converted into fatty acid ethyl esters and acetylated alcohols, which both can then be analyzed by gas chromatography (Section 24.3.1).

Cholesterol and Cholesterol Esters Cholesterol and cholesterol esters (Figure 24.10) play a major role as constituents of biomembranes (free cholesterol), but are also important for the pathogenesis of atherosclerosis. Cholesterol esters occur in lipoproteins of low (LDL) and high (HDL) density in the blood. Because of the major biological relevance of these neutral lipids several analytical methods have been developed to quantify their contents in biological samples. For quantification of serum cholesterol test kits are commercially available that differentiate between free and esterified cholesterol. With these kits, free serum cholesterol is oxidized by a bacterial cholesterol oxidase, yielding H_2O_2. This peroxide is subsequently reduced to H_2O in the presence of an oxidizable redox indicator, which changes color upon oxidation. Since cholesterol oxidase only reacts with free cholesterol, cholesterol esterase is added as helper enzyme. The difference between cholesterol esterase treated (total cholesterol) and untreated (free cholesterol) plasma quantifies the share of esterified cholesterol.

Besides the cholesterol oxidase method several chromatographic methods can be used for quantification of these neutral lipids. For instance, in a single HPLC run free and various esterified cholesterol subspecies can separately be quantified (Figure 24.11).

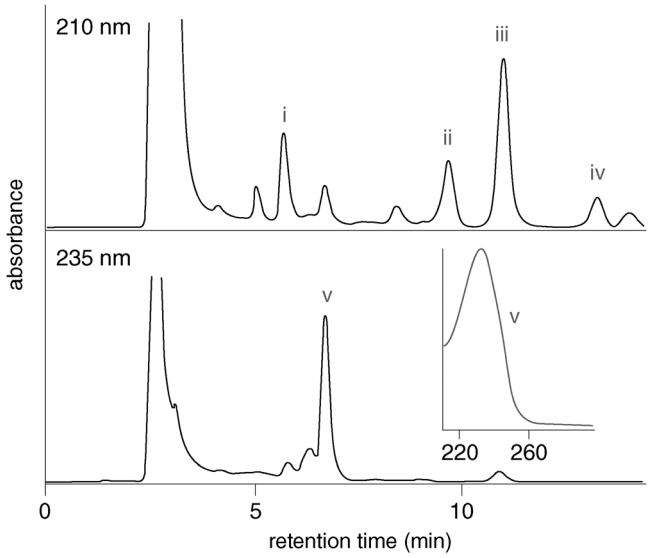
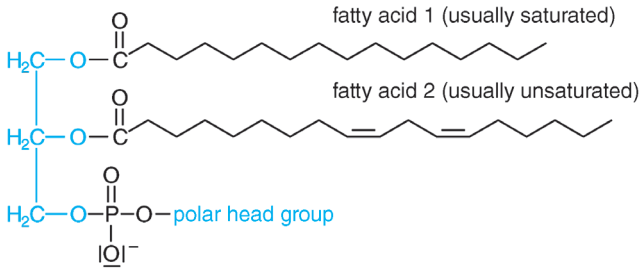


Figure 24.11 NP-HPLC analysis of free cholesterol and various cholesterol esters. A total lipid extract of human LDL was analyzed on a Nucleosil-50 column (250 mm × 5 mm, 7 µm particle size) using a mobile phase consisting of acetonitrile/2-propanol (75 : 25, by vol.) at 45 °C. Solvent flow: 1 ml min^{-1}. (i) Free cholesterol, (ii) cholesterol arachidonate, (iii) cholesterol linoleate, (iv) cholesterol oleate, and (v) cholesterol esters containing an oxidized linoleate residue. Inset: UV-spectrum of compounds eluted at (v). UV absorbance at 235 nm (oxidized cholesterol esters) and 205 nm (non-oxidized cholesterol derivatives) were recorded simultaneously.

24.4.4 Polar Ester Lipids

Phospholipids According to their structural characteristics phospholipids can be classified into different subfamilies (Figure 24.12). For analytical separation of the different phospholipid classes several thin-layer chromatographic systems (Section 24.3.1) have been developed. The major phospholipid classes of animal tissues can be separated on silica gel plates using solvent systems consisting of chloroform/methanol/water/acetic acid mixtures (Table 24.4). Since phosphatidylserines and phosphatidylinositols have similar polarities



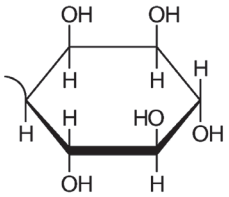
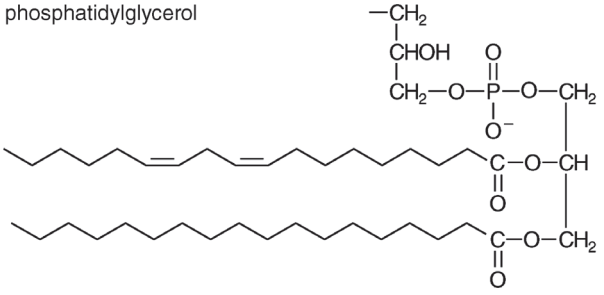
	polar head group	class of phospholipids
—	—H	phosphatidic acids
ethanolamine	$-CH_2-CH_2-\overset{+}{N}H_3$	phosphatidylethanolamines
choline	$-CH_2-CH_2-\overset{+}{N}(CH_3)_3$	phosphatidylcholines
serine	$-CH_2-CH(COO^-)-\overset{+}{N}H_3$	phosphatidylserines
inositol		phosphatidylinositols
phosphatidylglycerol		cardiolipins

Figure 24.12 Chemical structure of various glycerophospholipid classes. Glycerophospholipids consist of the trivalent alcohol glycerol, two fatty acids, a phosphate group, and a polar head group. The OH group at C1 of the glycerol backbone is usually esterified with a saturated fatty acid. At C2 of the glycerol moiety an unsaturated fatty acid is attached. The polar head group is linked via a phosphodiester bridge to C3 of the glycerol backbone. According to the chemistry of the polar head group glycerophospholipids can be classified into phosphatidylcholine, phosphatidylethanolamine, phosphatidylserine, and phosphatidylinositol. Cardiolipins form a special class of glycerophosphatides since their polar head group (glycerophosphatidic acid) involves additional fatty acids. These phospholipids mainly occur in the inner mitochondrial membrane and play an important role in the maintenance of the proton gradient (proton trap) over the inner mitochondrial membrane, which is formed by the respiratory chain. Another group of phospholipids, which function as important constituents of biomembranes, are sphingomyelins. They are typical components of lipid rafts and contain as structural component a sphingoid base instead of a glycerol moiety. The analytical methods used for sphingomyelins will be explained in detail in the sphingolipid subsection.

Table 24.4 Thin-layer chromatographic separation of different phospholipid classes. A lipid extract of rat liver mitochondria was separated on a silica gel plate with a solvent system consisting of chloroform/methanol/water/acetic acid (65 : 25 : 4, by vol.).

Phospholipid class	R_f value
Lysophospholipids	0.09
Phosphatidylserines, phosphatidylinositols	0.20
Sphingomyelins	0.30
Phosphatidylcholines	0.45
Phosphatidylethanolamines	0.70
Cardiolipins	0.81
Neutral lipids	0.92

they are hardly separated by one-dimensional thin-layer chromatography. Separation of this critical pair of analytes can be achieved by two-dimensional thin-layer chromatography. Employing high performance thin-layer chromatography (HPTLC) the separation of phospholipid classes was considerably improved.

For the separation of various phospholipid classes several HPLC systems have been developed. Coupling of HPLC with MS/MS recently became the method of choice in phospholipid analytics (phospholipidomics). This method is very sensitive and generates a large body of analytical data within a single chromatographic run. Figure 24.13 gives an example for a phospholipidome analysis.

To quantify the phospholipids in biological samples the phosphate content of washed total lipid extracts (Section 24.2) can be determined. For this purpose the solvent is evaporated and the remaining lipids are treated with a mixture of 10 N sulfuric acid and

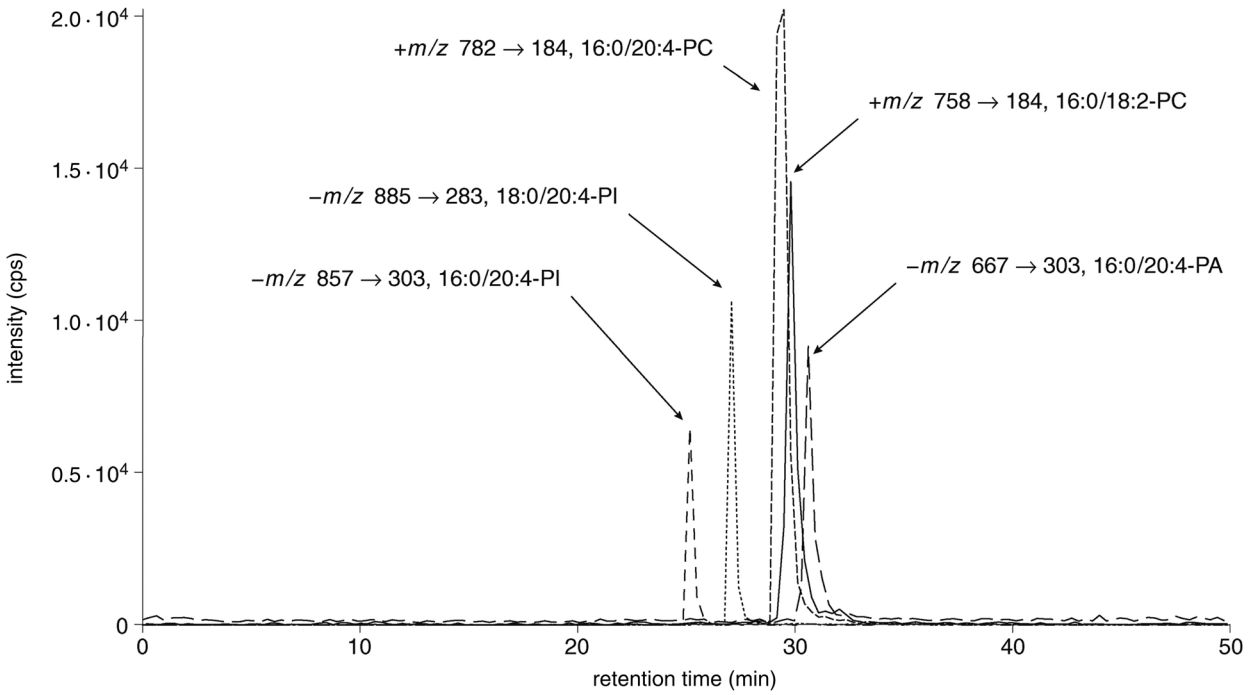
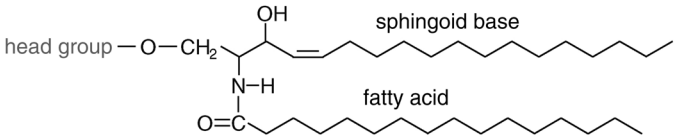


Figure 24.13 LC-MS/MS analysis of a mixture of phospholipid standards. The lipid mixture was analyzed on a Luna RP-HPLC column (150 mm × 2 mm, 3 µm particle size, Phenomenex, CA) using a linear gradient of two solvent systems A and B starting at 50% B and reaching 100% B in 10 min followed by isocratic elution for 30 min at 100% B. Solvent A: methanol/acetonitrile/water (60 : 20 : 20, by vol.) containing 1 mM ammonium acetate; solvent B: 100% methanol containing 1 mM ammonium acetate; flow rate 0,2 ml min^{-1}. As mass detector an AB Sciex 4000 Q-Trap system was employed. Source: data were kindly provided by Dr. V.D. O'Donnell, University of Cardiff.



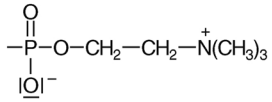
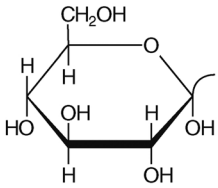
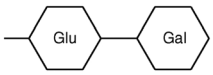
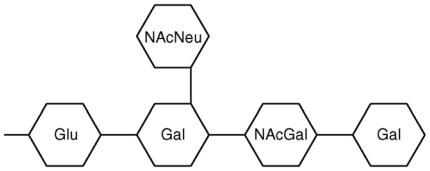
	polar head group	class of phospholipids
—	—H	ceramide
phosphocholine		sphingomyelin
glucose		glycosyl cerebroside
di- or trisaccharide		lactosylceramide
complex oligosaccharide		ganglioside

Figure 24.14 Chemical structure of different sphingolipids. The sphingoid base (sphingosine, sphinganine, hydroxysphinganine) forms the backbone of this lipid class and a saturated fatty acid is coupled to the sphingoid base via an amide bridge. Unsaturated fatty acids are rare components of glycosphingolipids. According to the chemical identity of their polar head groups (e.g., choline in sphingomyelin) sphingolipids can be separated into different subclasses.

60% perchloric acid for 30 min at 200 °C. The reaction products are reconstituted in 0.15 ml of distilled water and incubated with 0.8 ml of malachite green–ammonium molybdate solution for 10 min at room temperature. The resulting green complex can then be quantified at 660 nm. To avoid contamination of the samples care should be taken that prior to analysis all reaction tubes are rinsed with concentrated sulfuric acid and that phosphate-free water is always used.

Ion Exchange Chromatography, Section 10.4.7

Glycosphingolipids Beside glycerophospholipids, glycosphingolipids (Figure 24.14) form a second class of polar lipids. Neutral and simple glycolipids can be recovered from biological samples by Folch extraction (Section 24.2.1). However, gangliosides containing complex polysaccharide side chains partly remain in the water phase during Folch extraction. They can be isolated by solid phase extraction using C-18 cartridges (Section 24.2.2) and/or by anion exchange chromatography from the water phase of the Folch extracts. For complete structural elucidation of the glycosphingolipids complex analytical strategies must be applied to explore the composition, the sequence, and the anomeric configuration of the carbohydrate moiety as well as the chemical identity of the fatty acid component and the sphingoid base. Purified glycolipids can be fragmented into their structural components by treatment with methanolic HCl. The resulting methanolysis products (carbohydrates, fatty acid methyl esters, sphingoid bases) can then be further analyzed by chromatographic techniques. Mass spectrometry (Section 24.3.2) can be applied for analysis of unfragmented glycosphingolipids. By employing this method the carbohydrate moiety (composition and sequence) as well as the composition of the ceramide component can be explored. To separate the carbohydrates from the lipid components glycosphingolipids can be treated with ceramide gluconases and for sequencing of the carbohydrate moiety the lipids are treated with exoglucosidases. This enzyme catalyzes the stepwise liberation

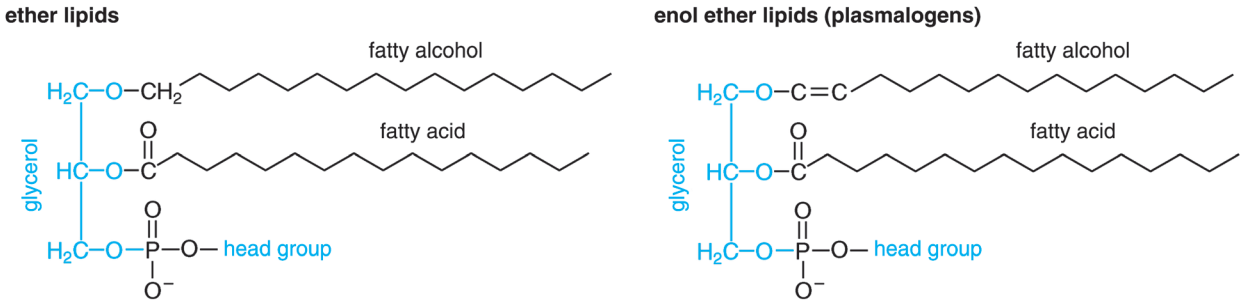


Figure 24.15 Chemical structure of ether lipids and plasmalogens. In ether lipids the glycerol moiety is connected with a long-chain fatty alcohol via an ether bridge. The other structural elements correspond to the glycerophospholipids. For enol ethers the fatty acyl chain at C1 of the glycerol moiety involves a double bond between the first and second carbon atom of the hydrocarbon chain. This enol ether bond confers special properties (lability against acids) to plasmalogens, which are utilized for analytical purposes.

of monosaccharides from the oligosaccharide chains and MS of the reaction mixture allows determination of its chemical identity.

Carbohydrate Analysis, Chapter 23

Unfragmented glycosphingolipids can also be analyzed by thin-layer chromatography (Section 24.3.1) on silica gel plates. When the plates are treated prior to analysis with boric acid the separation of different glycosphingolipids can be improved since boric acid interacts with vicinal OH groups of the carbohydrate moiety and, thus, impacts their retention behavior.

In addition to thin-layer chromatography several HPLC methods have been developed for the separation of glycosphingolipids. For this purpose silica gel columns are frequently used. Moreover, anion exchange chromatography can be applied.

Ether and Enol Ether Lipids Some animal tissues (heart muscle) are rich in ether lipids. The biological functions of these lipids in higher animals are not entirely clear but anti-oxidative functions have been suggested. In archaea, ether lipids are the major polar lipids in the cell envelope and their abundance is one of the major characteristics that separates this group of prokaryotes from bacteria. In these lipids the fatty acyl chain is coupled to the glycerol moiety via an ether bond (Figure 24.15). For plasmalogens the fatty acyl chain is coupled to C1 of the glycerol moiety via an enol ether bond whereas C2 of the glycerol is esterified with a fatty acid. The enol ether bond is acid-labile and can easily be cleaved at low pH. Acidic hydrolysis leads to the formation of a long-chain aldehyde, which can be analyzed after suitable derivatization by gas chromatography (Section 24.3.1). For quantification of the plasmalogen content in a complex lipid extract the phosphate content in the 1-lysophosphatide fraction following acid hydrolysis can be determined. However, this is a rather challenging task since it requires the precise preparative separation of the 1-lysophosphatide fraction from the other phospholipid classes by thin-layer chromatography or HPLC (Section 24.3.1). Since plasmalogens differ from structurally related ester phospholipid with respect to their molecular weight and their fragmentation behavior they can be analyzed directly without prior derivatization by electrospray ionization mass spectrometry (Section 24.3.2).

24.4.5 Lipid Hormones and Intracellular Signaling Molecules

Steroid Hormones For quantification of steroid hormones (Figure 24.16) in the blood immunoassays are frequently employed. Complete analytical systems are commercially available and the kits include detailed instructions for their use. The use of monoclonal antibodies guarantees high sensitivity and specificity. To avoid radioactive waste, enzyme immunoassays, fluorescence immunoassays, or chemiluminescence immunoassays are possible. Pre-purification or concentration of the analytes from body fluids are usually not required. More recently, mass spectrometric techniques (MALDI) have been developed to quantify most steroid hormones in the blood. For this purpose a drop of blood is placed on regular filter paper, which can subsequently be analyzed by MS.

Immunoassays, Section 5.3.3

Mass Spectrometry, Chapter 15

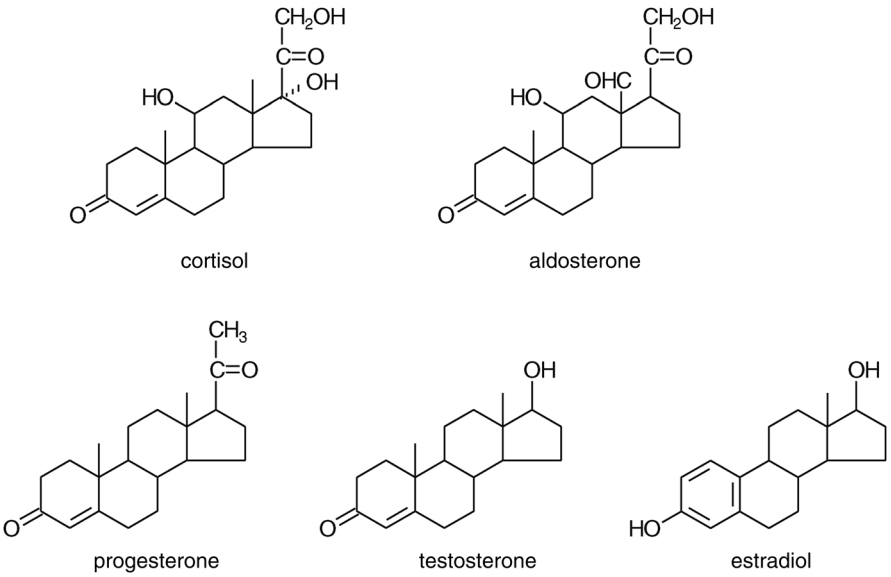


Figure 24.16 Chemical structure of selected steroid hormones. In the human body several steroid hormones are biosynthesized in various organs (adrenal gland, gonads, placenta, adipose tissue) using cholesterol as substrate. They are delivered into the blood and transported to target tissues. The synthetic routes involve several reaction intermediates, which also may exhibit biological activities.

Eicosanoids Eicosanoids are lipid hormones (Figure 24.17) that are biosynthesized from arachidonic acid ($\Delta^{5,8,11,14}$-all-*cis*-eicosatetraenoic acid) and exhibit pleiotropic biological activities (regulation of renal function, blood platelet aggregation, uterus contraction, etc.). Moreover, they have been implicated in pathophysiological processes such as allergy, inflammation, and thrombosis. They are biosynthesized via the arachidonic acid cascade (Figure 24.18) and can be classified according to their chemical structures into three different groups:

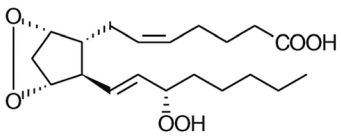
1. prostanoids (prostaglandins D_2, E_2, $F_{2\alpha}$, prostacyclin) and thromboxanes (TXA_2; TXB_2);
2. peptidoleukotrienes (leukotrienes C_4, D_4, E_4);
3. epoxy- and hydro(pero)xy fatty acids [(leukotrienes A_4, B_4, 5,6-epoxyeicosatrienoic acid, 12-hydro(pero)xyeicosatetraenoic acid (12-H(p)ETE)].

Historically, eicosanoids were quantified in biological lipid extracts by bioassays employing different cell (blood platelets) or organ (arteries) preparations. Later on, radio-thin-layer chromatography (Section 24.3.1) was frequently used.

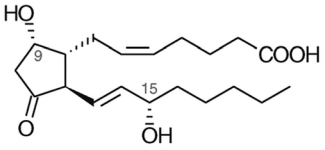
Today thin-layer chromatography is not longer employed for eicosanoid analysis since HPLC techniques have become more important. Unfortunately, unlike leukotrienes and hydroxy fatty acids, prostaglandins do not carry strong UV/Vis chromophores and, thus, sensitive UV-detection is hardly possible. In contrast, hydro(pero)xy fatty acids and leukotrienes contain conjugated double bonds and thus can easily be detected at 235 nm (conjugated dienes), 270 nm (conjugated trienes), and 300 nm (conjugated tetraenes). Using diode array photo-detection the chromatograms can be followed at different wavelengths and UV-spectra of all analytes can be extracted. Thus, the compounds of interest can be identified by two independent parameters (retention time, UV spectrum). For epoxy- and hydro(pero)xy fatty acids reverse phase HPLC on C_{18}-columns with methanol/water mixtures as mobile phases is frequently employed. For special purposes (separation of positional isomers, enantiomer analysis) normal phase or chiral phase HPLC are used. Peptidoleukotrienes are also analyzed on reverse phase HPLC columns.

In recent years several LC-MS/MS systems have been developed to analyze the eicosanome as part of the cellular lipidome (Section 24.6). With these methods more than 150 different eicosanoids and corresponding derivatives of other polyenoic fatty acids can be analyzed within a single analytical run (Figure 24.19). Although these systems are very sensitive and provide a large body of information there are some problems associated with these techniques. The more compounds that are analyzed within a single chromatographic run the higher is the likelihood for the occurrence of critical pairs of analytes, which can hardly be separated from each other. Moreover, conventional chromatographic techniques are not suitable for enantiomer separation. Since the mass spectra of enantiomers are identical regular LC-MS/MS systems cannot be used

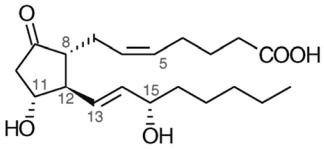
prostaglandines



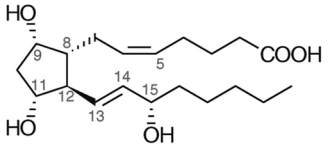
prostaglandin G_2



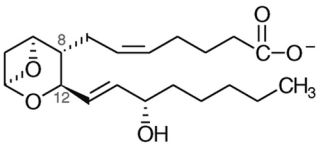
prostaglandin D_2



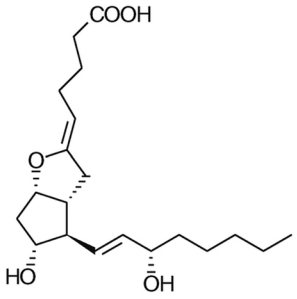
prostaglandin E_2



prostaglandin $F_{2\alpha}$

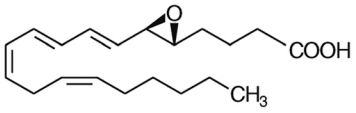


thromboxane A_2

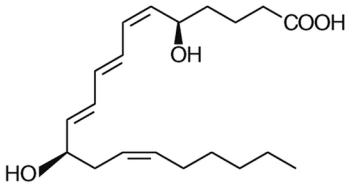


prostaglandin I_2

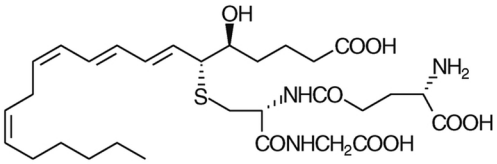
leucotrienes



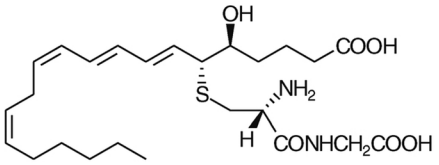
leukotriene A_4



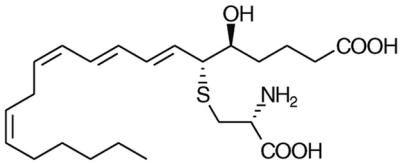
leukotriene B_4



leukotriene C_4

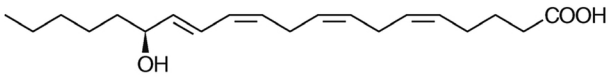


leukotriene D_4

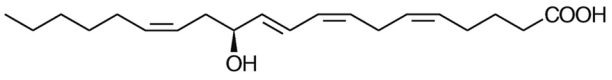


leukotriene E_4

hydroxy fatty acids



15-hydroxyeicosatetraenoic acid (15-HETE)



12-hydroxyeicosatetraenoic acid (12-HETE)

Figure 24.17 Chemical structure of selected eicosanoids. Eicosanoids are oxygenation products of arachidonic acid (Figure 24.18). Prostaglandins as well as epoxy- and hydro(pero)xy fatty acids are oxygenation products of arachidonic acid or other polyunsaturated fatty acids but do not contain other structural elements. In contrast, for leukotriene biosynthesis glutathione is needed in addition to arachidonic acid. In the biosynthetic cascade of prostaglandins arachidonic acid is initially oxygenated to the cyclic endoperoxide PG G_2, which is subsequently transformed by different enzymes into other prostaglandins, thromboxanes, and prostacyclins. During leukotriene biosynthesis the primary arachidonic acid oxygenation product (LT A_4) is either hydrolyzed enzymatically (LT A_4 hydrolase) to give LT B_4 or conjugated with glutathione to yield LT C_4. Sequential cleavage of the glutathione moiety (glutamate, glycine) by γ-glutamyl-transpeptidases and/or other peptidases leads to the formation of LT D_4 and LT E_4, which also exhibit bioactivity. To avoid confusion products of the third route of the arachidonic acid cascade, the cytochrome P450 pathway, are not shown here.

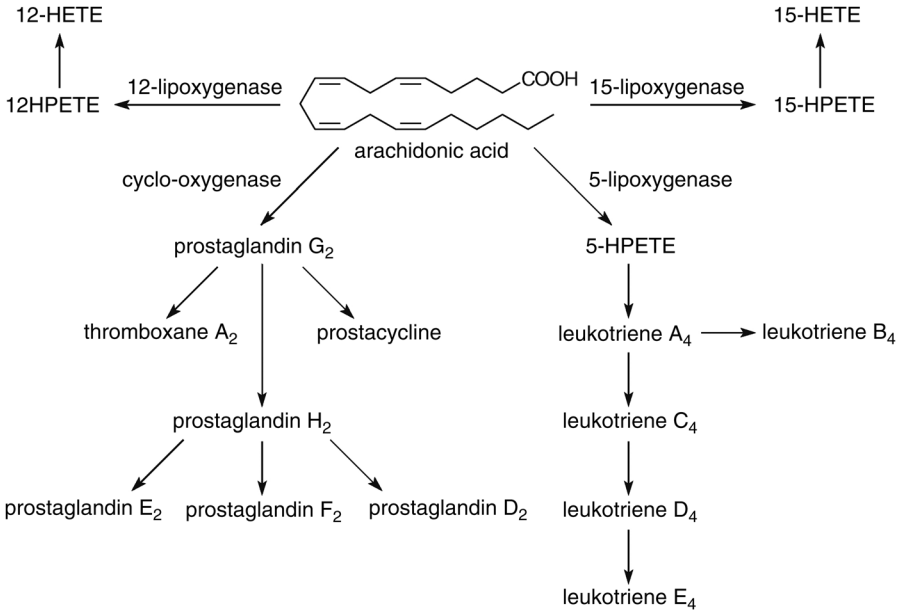


Figure 24.18 Biosynthetic scheme of eicosanoid formation (arachidonic acid cascade). The chemical structures of the different compounds are given in Figure 24.17.

for enantiomer analysis. Online derivatization of the enantiomers to diastereomeric derivatives using chiral resolving agents or the use of chiral stationary phases may solve this problem in the future. However, for the time being such systems have not gained widespread acceptance.

Alongside HPLC, capillary gas chromatography is an important element in the analytical hierarchy of eicosanoids; it is usually combined with mass spectrometry. Prostaglandins and hydroxy fatty acids are usually analyzed on apolar or medium polar stationary phases following suitable derivatization (esterification of the carboxylic group, silylation of the hydroxy groups). When coupled with mass spectrometry retention times and typical fragmentation patterns indicate the structure of the analytes.

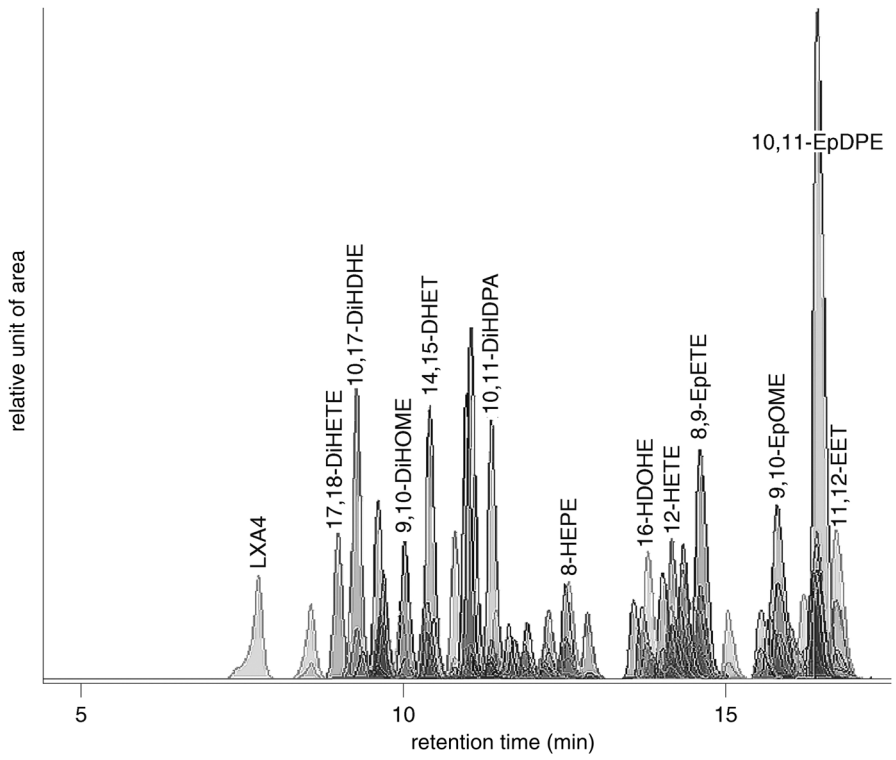


Figure 24.19 LC-MS analysis of eicosanoids. A mixture of synthetic standard eicosanoids was analyzed by LC-MS/MS at 40 °C on a Zorbax C_{18}-Eclipse plus column (2.1 mm × 150 mm, 3.5 µm particle size) using a binary solvent gradient (solvent A: 100% acetonitrile, solvent B: 10 mM ammonium acetate). Gradient: between 0 and 1 min – 95% solvent B, between 1 and 2 min from 95% B to 70% B, between 2 and 18.7 min from 70% B to 33% B, between 18.7 and 25 min from 33% B to 10% B; flow rate 0.4 ml min^{-1}. Analytical configuration: Agilent 1200SL HPLC-system coupled with an Agilent 6460 triple-quad-mass spectrometer (electrospray ionization). The following eicosanoids were injected for analysis (elution order): LXA4, Resolvin D1, LTB5, 17,18-DiHETE, 10,17-DiHDHE, 14,15-DiHETE, LTB4-D4, LTB4, 12,13-DiHome-D4, 11,12-DiHETE, 12,13-DiHome, 8,9-DiHETE, 9,10-DiHome, 14,15-DHET-D11, 19,20-DiHDPA, 14,15-DHET, 5,6-DiHETE, 16,17-DiHDPA, 11,12-DHET, 13,14-DiHDPA, 13,14-DiHDPA, 19-HEPE, 20-HEPE, 10,11-DiHDPA, 8,9-DHET, 18-HEPE, 7,8-DiHDPA, 5,6-DHET, 19-HETE, 20-HETE-d6, 20-HETE, 15-HEPE, 9-HEPE, 8-HEPE, 12-HEPE, 21-HDoHE, 5-HEPE, 22-HDoHE, 15-HETE-d8, 15-HETE, 16-HDoHE, 17-HDoHE, 13-HDoHE, 17,18-EpETE, 14-HDoHE, 10-HDoHE, 11-HETE, 8-HDoHE, 11-HDoHE, 7-HDoHE, 8-HETE, 9-HETE, 12-HETE, 5,6-EpETE, 5-HETE, 11,12-EpETE, 14,15-EpETE, 8,9-EpETE, 4-HDoHE, 12,13-EpOME-D4, 12,13-EpOME, 9,10-EpOME, 19,20-EpDPE, 14,15-EET-d8, 14,15-EET, 5,6-EET, 16,17-EpDPE, 10,11-EpDPE, 13,14-EpDPE, 7,8-EpDPE, 11,12-EET, 8,9-EET. Source: data were kindly provided by Dr. M. Rothe LIPIDOMIX GmbH, Berliner Allee 261-262, 13088 Berlin.

In recent years immunoassays have been developed for various eicosanoids and kits, which include standards and detailed user instructions, are commercially available. Although some of these kits work well with biological fluids, such as plasma, synovial fluid, or urine, pre-purification and concentration of the analytes are sometimes required. A major problem of such immunoassays might be the possible cross-reactivity of the employed antibodies with other components of the extracts. Although cross-reactivity with selected structurally related eicosanoids has been excluded it remains uncertain whether unknown derivatives might cross-react. Thus, analytical data obtained with immunoassays should be confirmed at least once by a physicochemical method (LC-MS/MS).

Diacyl Glycerols Diacyl glycerols (DAGs) are important intracellular signaling molecules, which activate protein kinase C and, thus, contribute to regulation of intracellular metabolism. Agonist binding at certain cell surface receptors activates phospholipase C (Figure 24.2), which hydrolyzes phosphorylated phosphatidylinositols of the cell membrane leading to the formation of diacylglycerols and inositol-3-phosphate (IP3). IP3 is water soluble, diffuses into the cytosol, and functions as Ca-ionophore. It binds to IP3-dependent Ca-ion channels located in the membrane of the endoplasmic reticulum and triggers Ca release into the cytosol. Because of its hydrophobic character DAG remains in the plasma membrane and activates membrane bound protein kinase C, which in turn phosphorylates different metabolic enzymes and constituents of the cytoskeleton. DAGs can be analyzed by thin-layer chromatography on silica plates (Section 24.3.1) or by SP-HPLC using solvent systems consisting of *n*-hexane/2-propanol (100 : 2, by vol.) Since they occur in cells and tissues at rather low concentrations and since they do not carry UV/Vis-chromophores detection of DAGs is problematic. To circumvent these problems mass sensitive detectors may be used. Alternatively, cells can be pre-labeled with radioactive fatty acids. When the cells are subsequently stimulated with receptor agonists that activate phospholipase C radioactively labeled DAG are formed, which can be quantified by radio-thin-layer chromatography.

Sphingosine Derivatives Sphingosines are long-chain amino alcohols (Figure 24.20), which for a long time have been considered mainly as structural constituents of membrane lipids. They have been identified as major lipid components of the skin and have been implicated in the water barrier of the epidermis. However, more recently, sphingolipids have been identified as signaling molecules that play a role in growth regulation, cell differentiation, and apoptosis. Ceramides exhibit pro-apoptotic properties. In contrast, sphingosine-1-phosphate (S1P) is an anti-apoptotic agent and thus constitutes a physiological counter player of ceramides. In addition, S1P functions as intracellular second messenger but can also be released into the extracellular compartment. Here, it functions as lipid hormone, binding at S1P receptors (S1P1–S1P5) of target cells. Receptor–ligand interaction induces intracellular liberation of cAMP, IP3, and Ca^{2+}, which modifies the cellular gene expression profile and induces rearrangements of the cytoskeleton. In peripheral blood S1P is present at fairly high concentrations (up to 1 μM). It has been suggested as an important regulator of neoangiogenesis in solid tumors and, thus, S1P receptor antagonists and S1P synthesis inhibitors have been developed as antitumor drugs. RP-HPLC has been the method

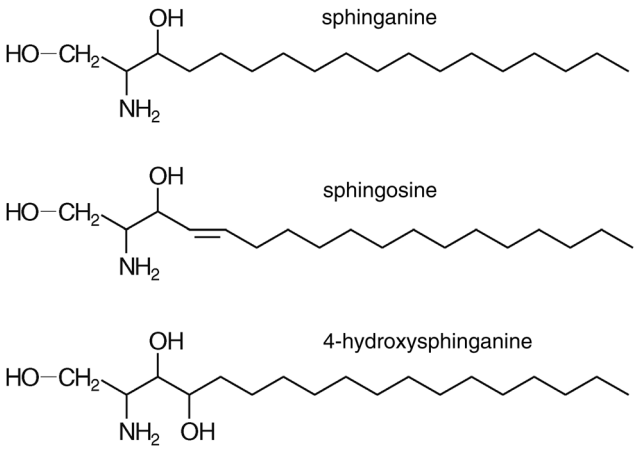


Figure 24.20 Chemical structures of biologically relevant sphingoid bases. Sphingoid bases are long-chain amino alcohols. In biological samples three major sphingoid bases occur, which consist of 18 carbon atoms.

of choice for analyses of sphingosine derivatives. Since these compounds do not carry suitable UV/Vis chromophores they are usually derivatized with phthalate aldehydes to introduce a strong fluorophore. Today several analytical LC-MS/MS methods are available that do not require pre-analytical derivatization.

24.5 Lipid Vitamins

Vitamins are organic compounds that are needed for proper functionality of an organism. They are required in small amounts but cannot be endogenously synthesized in sufficient quantities. Thus, they must be taken up with the diet. In this regard the term vitamin is conditional upon the nutritional circumstances and the particular organism. For humans, water- and lipid-soluble vitamins can be differentiated but for this chapter we focus on lipid-soluble vitamins only. The most important lipid-soluble human vitamins are vitamins A, D, E, and K. Some lipid vitamins can partly be synthesized in the human body particularly by gut microbes. In some cases pro-vitamins are taken up and are then converted into the bioactive vitamins. Vitamin K is synthesized in large amounts by gut microbes and alterations of the enteral microbiome as a consequence of chronic inflammatory bowel diseases may lead to defective endogenous vitamin K synthesis. For diagnosis of hypovitaminoses, which frequently occur in developing countries as a consequence of general malnutrition, sensitive analytical methods are required to quantify the blood level of active lipid vitamins and their inactive synthetic precursors. Because of their high evaporation temperatures and their thermal lability gas chromatography is not well suited for this purpose. However, several rather simple HPLC systems have been developed, which are easy to use even with standard HPLC configurations (binary gradients with UV/Vis or fluorescence detection).

Vitamin A and Carotenoids Vitamin A (retinol) is absorbed in the gut either as biologically active vitamin or as β-carotene, which functions as pro-vitamin (Figure 24.21). HPLC is the method of choice for retinol quantification in biological fluids. Since retinol contains a strong fluorophore (excitation at 330 nm, emission at 480 nm) it can easily be detected in the column effluent. Since fluorimetric measurements are very sensitive and non-invasive it is possible to reliably quantify retinol concentrations in the blood. If there is no online fluorescence detector UV absorbance at 325 nm can be quantified but here the detection limit is at least one order of magnitude higher (low sensitivity).

Carotenoids, the pro-vitamins of retinol, are much more hydrophobic than retinol itself (Figure 24.21) and, thus, stronger eluting solvent systems are required for RP-HPLC. Since the spectral properties and the molar extinction coefficients are different for different carotenoids (Table 24.5) multi-wavelength detection or the use of diode array detectors is strongly advised.

Vitamin D Vitamin D_3 (dihydroxy cholecalciferol, calcitriol) can be completely synthesized in the human body and thus, by definition, does not constitute a classical vitamin. In fact, since it can completely be synthesized from cholesterol, since it is transported in the blood to its target cells, and since it exhibits signaling properties it may rather be considered a hormone. The major bioactivity of vitamin D_3 is regulation of extracellular calcium homeostasis but it has also been

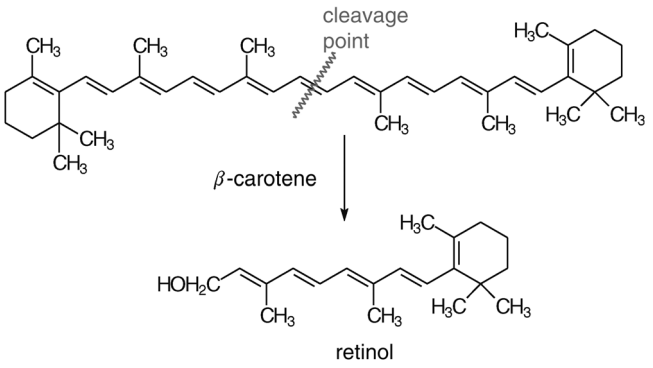


Figure 24.21 Chemical structures of β-carotene and retinol (vitamin A). Cleavage of the indicated double bond converts the pro-vitamin β-carotene into biologically active retinol (vitamin A), which is required for signal transduction in the retina and as regulator of cell differentiation and maturation.

Table 24.5 UV-properties of selected carotenoids and of α-tocopherol.

Carotenoid	λ_{max} (nm)	Molar extinction coefficient (M^{-1} cm^{-1})
Retinol	325	53 000
Retinyl stearate	325	53 000
Lutein	454	144 000
Cryptoxanthin	454	131 000
Lycopene	474	185 000
α-Carotene	447	146 000
β-Carotene	450	149 000
γ-Carotene	454	137 000
α-Tocopherol	293	4070

implicated in cell maturation and differentiation. Biosynthesis of vitamin D_3 starts in the liver by dehydration of cholesterol to 7-dehydrocholesterol (Figure 24.22). This intermediate is subsequently transported to the skin. Here the B-ring of the sterane backbone is oxidatively cleaved in a reaction that requires UV light. The resulting cholecalciferol is then converted in the liver into 1-hydroxy-cholecalciferol and finally in the kidney into 1,25-dihydroxy-cholecalciferol, the bioactive 1,25-calcitriol. When final hydroxylation in the kidney oxidizes the 24th carbon atom of the cholesterol backbone instead of C1 biologically inactive 24,25-dihydroxy-cholecalciferol (24,25-calcitriol) is formed. The major bioactivity of 1,25-calcitriol is upregulation of the calcium concentration in the blood plasma by impacting enteral calcium resorption, urinary calcium excretion, and calcium mobilization from the bones.

For quantitative analysis of various vitamin D derivatives several HPLC systems have been developed. Because of the conjugated double bonds found in the naturally occurring vitamin D derivatives UV-detection at 254 nm is possible.

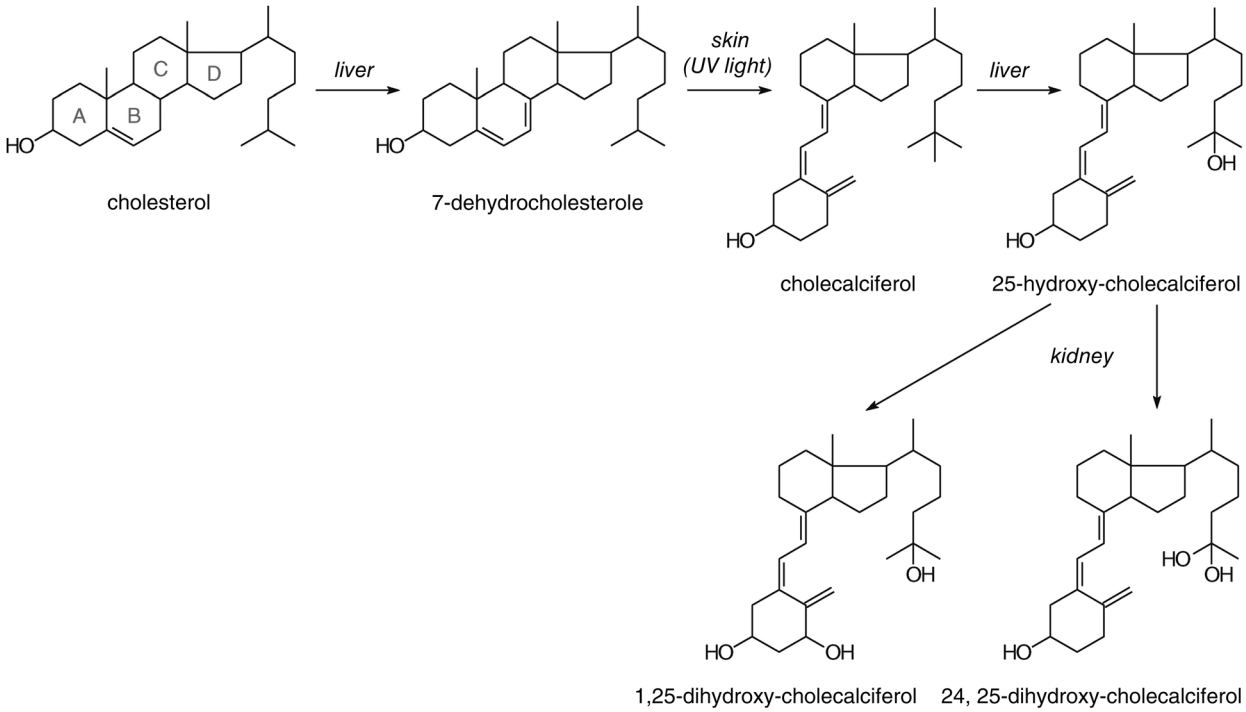


Figure 24.22 Biosynthesis of 1,25-dihydroxy-cholecalciferol (calcitriol, vitamin D_3). Three different organs (liver, skin, kidney) of the human body are involved in the biosynthesis of vitamin D_3.

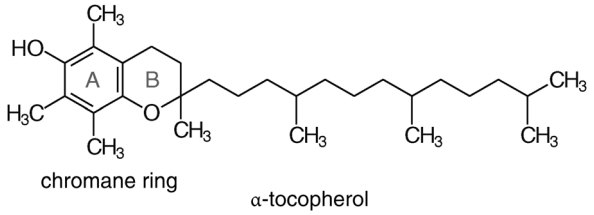


Figure 24.23 Structure of α-tocopherol. The B-ring of α-tocopherol can be cleaved, forming α-tocopherol hydroquinone. This intermediate is then converted into tocoquinone by intermediate formation of the hydrosemiquinone radical. The delivered electrons can be picked up by free radicals in the environment, which explains the radical scavenging properties of α-tocopherol.

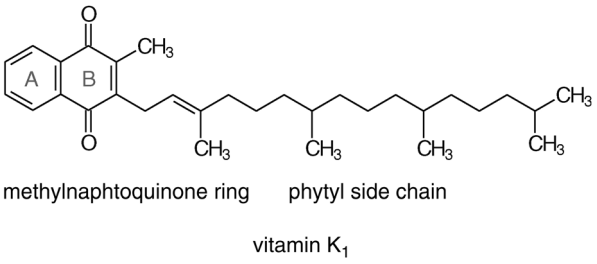


Figure 24.24 Structure of vitamin K_1. Vitamin K_2 carries a difarnesyl residue consisting of six isoprenoid units instead of a phytyl side chain.

Vitamin E Tocopherols (Figure 24.23) consist of a chromane ring and an isoprenoid side chain. The different types of tocopherols (α-tocopherol, γ-tocopherol) structurally differ from each other by the number and position of the methyl groups bound at the chromane ring. Tocopherols protect lipids, carbohydrate, nucleotides, and amino acids from oxidation (anti-oxidant function) and are frequently named radical scavengers. In biological systems tocopherols are mainly found in the phospholipid bilayer of biomembranes but also in lipoproteins. Here they protect the unsaturated fatty acids of these complex lipid–protein assemblies from radical induced oxidation.

Tocopherols can easily be quantified by RP-HPLC without prior derivatization. Since they carry an aromatic ring the chromatograms can be followed by measuring the UV absorbance at 290 nm. However, because of the low molar extinction coefficient (Table 24.5) UV detection is not very sensitive. Alternatively, fluorescence measurements (excitation at 292 nm, emission at 325 nm) can be carried out if an online fluorescence detector is available.

Vitamin K The K vitamins (phylloquinones) are derivatives of menadione, which does not occur in nature (Figure 24.24). They function as cofactors for post-translational modification of proteins and are essential for γ-carboxylation of coagulation factors and of calcium binding proteins. K-hypovitaminoses might lead to prolonged bleeding times and to impairment of bone calcification.

Since the side chains of various vitamin K isoforms are different in length the hydrophobic properties of the isoforms are very different. This is of relevance for the analytical systems separating the different isoforms on the basis of their polarity. In RP-HPLC, isocratic elution is only possible for selected vitamin K isoforms. For complete separation of the major naturally occurring vitamin K isoform gradient elutions are required. To sensitively detect the analytes in the column effluent one can take advantage of the intrinsic fluorescence properties of the analytes (excitation at 320 nm, emission at 430 nm).

24.6 Lipidome Analysis

The lipidome of a living cell consists of all lipid specimens that are present in the cell at the time of analysis. Although most cellular lipids consist of a relatively low number of key constituents (fatty acids, glycerol, sphingoid bases, polar head groups, isoprenoids) human cells and tissues contain thousands of structurally distinct lipid species. This structural diversity is the result of multiple biosynthetic pathways and is based on the combinatorial multiplicity of the key structural elements. As a consequence of the rapid development of genomic, transcriptomic, and proteomic research there have been multiple attempts to quantify the cellular lipidome as

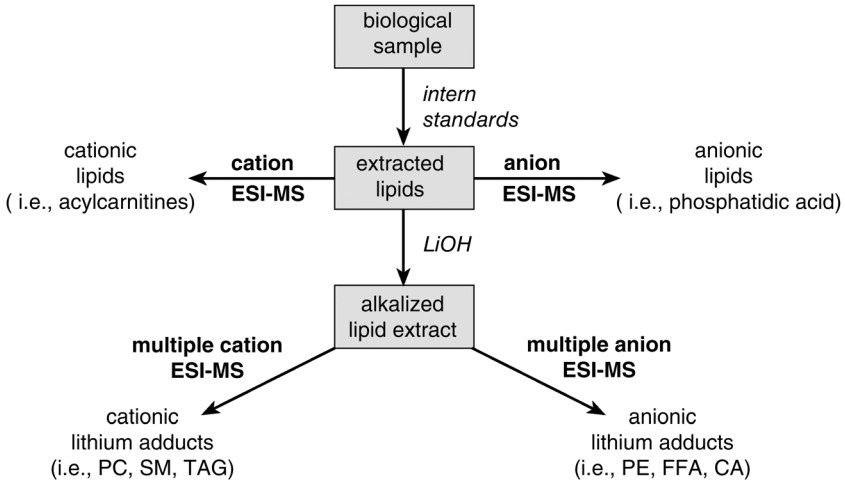


Figure 24.25 Experimental strategy for mass spectrometric analysis of lipidomes of biological samples. All of the lipids were extracted from weakly acidified biological samples and were subsequently analyzed by ESI-MS. In a second step the extracts were treated with LiOH and the alkalized extracts were analyzed again.

part of the cellular metabolome. The initial attempts to quantify cellular lipidomes were based on the combination of several chromatographic techniques. However, even after mass sensitive detectors were introduced the sensitivity of these initial approaches was hardly sufficient to quantify such lipid mediators, which only occur in pg amounts in biological systems. Moreover, to reliably analyze lipids, in particular neutral lipids, by mass spectrometry reliable ionization methods are needed and this remained a problem for many years. The introduction of electrospray mass spectrometry (ESI-MS, Section 16.2) partly solved this problem. Today it is possible to analyze complex lipid extracts of biological samples by online combination of HPLC with ESI-MS/MS. In a single analytical run quantitative analytical data can be obtained for hundreds of different lipid analytes. Major pre-purification of the lipid extracts might improve the performance but it is usually not required.

Figure 24.25 summarizes the experimental strategy for lipidome analysis of a cellular lipid extract. Because of the high sensitivity of ESI-MS 10^6 cells or 10 mg of tissue (wet weight) are sufficient to quantify the lipidome. Initially, the total lipids are extracted from the biological samples. For this purpose the tissue is homogenized in organic solvents and suitable external deuterated standards are added to normalize the recovery of the extraction procedure. Then the lipid extracts can be analyzed directly by ESI-MS. Anionic lipids (cardiolipins, phosphatidylglycerols, phosphatidylinositols, phosphatidylserines, phosphatidic acids, and sulfatides) can be analyzed directly and this is also the case for acyl carnitines. Lipid classes with less pronounced charges (phosphatidylcholines, phosphatidylethanolamines, sphingomyelins, ceramides, galactocerebrosides, etc.) are less well ionized by ESI. To improve analytical quality the lipid extracts are treated with LiOH (50 nmol per mg cellular protein). This treatment provides negative counter-ions so that the ion pairs can be analyzed in the positive ionization mode. Simultaneously, the uncharged phosphatidylethanolamines are charged and can be analyzed in the negative ionization mode. Reliable analysis of strongly apolar neutral lipid (triacyl glycerols, cholesterol esters) still remains a problem. Fortunately, triacyl glycerols also form positively charged lithium complexes and, thus, can be analyzed by mass spectrometry (Figure 24.8). Lipidome analysis should always be carried out with diluted lipid solutions. When concentrated lipid extracts are used intense hydrophobic interactions between the analytes may impair the ionization process and, thus, the lipid content is underestimated.

ESI-MS, Section 15.1.2

The advantages of ESI-MS (Figure 24.26) based lipidome analysis can be summarized as follows:

1. simultaneous quantitative analysis of individual lipid species without prior pre-purification and derivatization;
2. comprehensive databases obtained in a single analytic run;
3. high sensitivity and good signal-to-noise ratio;
4. linear dependence of signal intensity on analyte concentration over a large concentration range (up to four orders of magnitude);
5. high reproducibility.

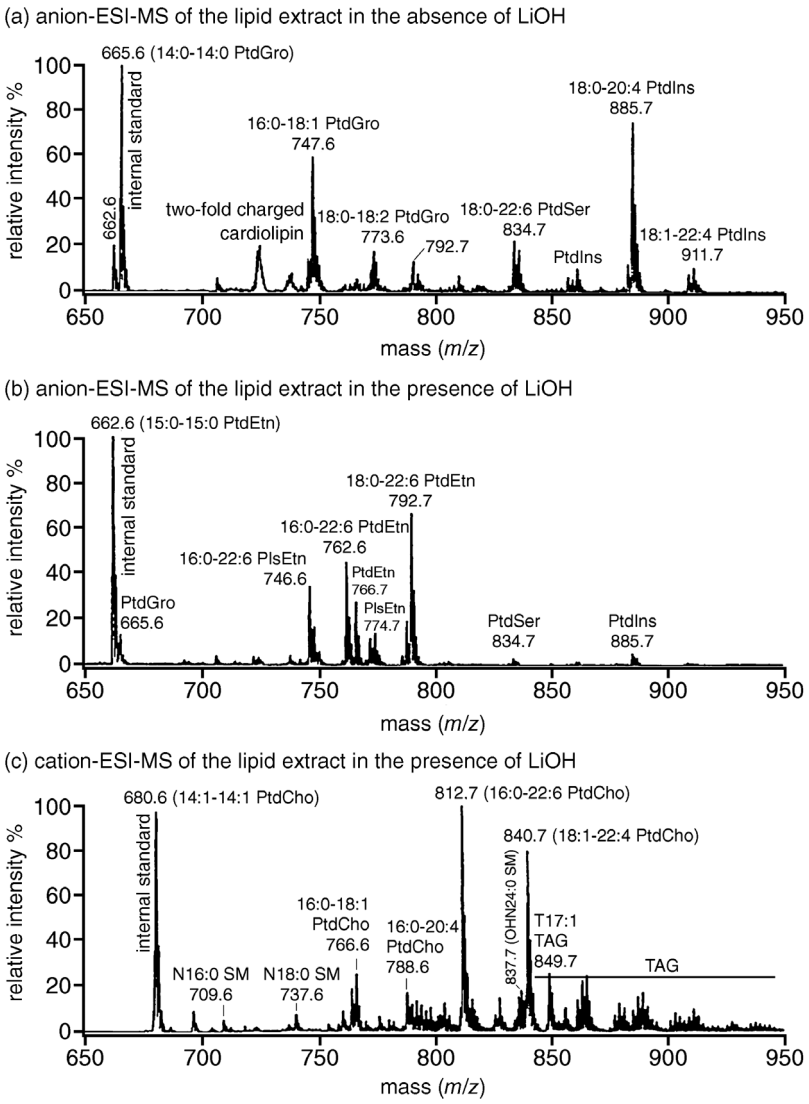


Figure 24.26 Electrospray ionization mass spectrometry of a mouse heart total lipid extract. The lipids were extracted from a homogenate of a mouse heart using the Bligh/Dyer method. The identity of the different lipid subspecies was confirmed by tandem mass spectrometry. (a) Anion-ESI-MS of the lipid extract in the absence of LiOH; (b) anion-ESI-MS of the lipid extract in the presence of LiOH; (c) cation-ESI-MS of the lipid extract in the presence of LiOH.

As disadvantages the uncontrolled ionization of certain neutral lipid classes and the high costs of the required analytical devices as well as of the deuterated external quantification standards may be discussed. Moreover, running such devices is more like an art and requires profound methodological experience. It is hardly possible to run such complex devices on top of other tasks.

24.7 Perspectives

Further development of analytical techniques will improve lipid analysis but in the upcoming years there will be no major groundbreaking innovations in the field of lipid analysis. Major advancements will be achieved with respect to improvement and simplification of existing analytical methods and devices. For the time being the high costs of LC-MS/MS devices and their technical complexity prevent widespread application of these methods. Lipidome analyses are usually carried out in specialized analytical centers so that intense collaborations between experimenters and analysts are required. It is important that all collaborators are involved in planning the experiments since methodological details of the workup procedure, such as selection of the most suitable extraction method or definition of extraction conditions (acidic or

alkaline extraction), may be important for the quality of the analytical data. Although it is hardly possible to reliably predict future development in the field of lipid analysis several trends might be discussed:

1. Miniaturization, simplification, and increased sensitivity of existing analytical methods: To make the analytical methods more user-friendly small compact analytical devices (tabletop devices) are likely to be constructed and the analytical procedures will be simplified. In current lipid research an increasing number of lipid signaling molecules (hormones, intracellular mediators) have been identified and many of them are found in biological samples only at subnanomolar concentration. To reliably quantify such low lipid concentrations more sensitive analytical methods need to be developed. *Narrow bore* HPLC that employs miniaturized HPLC columns and solvent flows in the µl range is likely to become more popular in lipid research. However, this technique requires high precision pumps and supersensitive mass detectors. Unfortunately, the currently existing HPLC systems can hardly be adapted to the narrow bore technology.
2. Serial and non-serial online coupling of various analytical methods: Although LC-MS/MS is currently the method of choice in lipidome analytics, online coupling of LC with other spectroscopic methods such as Fourier transform infrared spectroscopy (FTIR) or nuclear magnetic resonance spectroscopy (NMR) has already been worked out. Since NMR is one of the most powerful analytical methods LC-NMR online coupling provides important structural information on LC-separated individual lipids. Unfortunately, to obtain high quality analytical data at least µg amounts of analytes are required. In contrast, ng and in some cases even pg quantities of analytes are sufficient for MS detection. In addition to serial online coupling of non-invasive analytical methods, such as UV/Vis- or fluorescence spectroscopy, non-serial coupling approaches of invasive analytical methods are likely to be worked out. Since the chemical structures of the analytes are modified when invasive analytical methods, such as MS/MS, are applied serial online coupling of invasive methods is not meaningful. However, such methods can be combined with each other when the effluent of a chromatographic column is split (non-serial coupling) so that the separated analytes can further be explored in parallel using different invasive techniques.
3. Functional lipidomics: Using the currently available LC-MS/MS techniques it is possible to structurally characterize the lipidome of a biological sample under different conditions. However, while having such comprehensive structural information it still remains a problem to make conclusions concerning functional changes. Although we know that an increase in the cellular DAG content activates protein kinase C in a large number of experimental systems, the extent of activation is likely to be influenced by the presence or absence of other lipid and non-lipid mediators (mediator context). In other words, the biological effect of a single lipid mediator depends on the complex mixture of other signaling molecules and quantification of complex mediator profiles is more likely to mirror the functional alterations. Thus, quantification of cellular lipidomes provides more reliable data with which to conclude or even predict functional alterations than quantification of a single lipid mediator. To correlate the composition of cellular lipidomes with functional cell parameters large databases are required, which involve both the structural composition of the lipidomes as well as cell functional parameters. Unfortunately, such comprehensive databases are rarely available at present.
4. *In situ* mass spectrometry: Mass spectrometry is a powerful analytical method, which currently is mainly employed *in vitro*. However, *in situ* mass spectrometry can be applied to characterize the lipidome of fixed cell preparations or microscopic cross-sections. For this purpose a laser beam liberates a mixture of metabolites from certain areas of a microscopic cross-section or from an immobilized cell, which can then be analyzed by MS/MS. In this sense *in situ* MS is a variant of matrix-assisted laser desorption ionization mass spectrometry (MALDI). Although this method is very appealing there are two major problems with it: (i) The intensity of the laser desorption strongly depends on the chemistry of the matrix, which is clearly different when using different cell preparations. (ii) Standardization and exact quantification is virtually impossible since external quantification standards (deuterated analytes) can hardly be used.

Further Reading

Akoh, C.C. and Min, D.B. (2008) *Food Lipids*, CRC Press, Boca Raton, FL.

Bielawski, J. *et al*. (2010) Sphingolipid analysis by high performance liquid chromatography-tandem mass spectrometry (HPLC-MS/MS). *Adv. Exp. Med. Biol.*, **688**, 46–59.

Bligh, E.G. and Dyer, W.J. (1959) A rapid method of total lipid extraction and purification. *Can. J. Biochem. Physiol.*, **37**, 911–917.

Bollinger, J.G. *et al*. (2010) Improved sensitivity mass spectrometric detection of eicosanoids by charge reversal derivatization. *Anal. Chem.*, **82**, 6790–6796.

Chen, Y. *et al*. (2010) Imaging MALDI mass spectrometry of sphingolipids using an oscillating capillary nebulizer matrix application system. *Methods Mol. Biol.*, **656**, 131–416.

Christie, W.W. and Han, X. (2010) *Lipid Analysis – Isolation, Separation, Identification and Lipidomic Analysis*, Oily Press.

Deems, R. *et al*. (2007) Detection and quantitation of eicosanoids via high performance liquid chromatography-electrospray ionization-mass spectrometry. *Methods Enzymol.*, **432**, 59–82.

Fahy, E. *et al*. (2005) Comprehensive classification system for lipids. *J. Lipid Res.*, **46**, 839–861.

Fahy, E. *et al*. (2009) Update of the LIPID MAPS comprehensive classification system for lipids. *J. Lipid Res.*, **50**, S9–14.

Garrett, T.A. *et al*. (2007) Analysis of ubiquinones, dolichols, and dolichol diphosphate-oligosaccharides by liquid chromatography-electrospray ionization-mass spectrometry. *Methods Enzymol.*, **432**, 117–143.

Gunstone, F.D., Harwood, J.L., and Dijkstra, A.J. (2007) *The Lipid Handbook*, CRC Press, Boca Raton, FL.

Han, X., Yang, K., Cheng, H., Fikes, K.N., and Gross, R.W. (2005) Shotgun lipidomics of phosphatidylethanolamine containing lipids in biological samples after one-step in situ derivatization. *J. Lipid Res.*, **46**, 1548–1560.

Kasumov, T. *et al*. (2010) Quantification of ceramide species in biological samples by liquid chromatography electrospray ionization tandem mass spectrometry. *Anal. Biochem.* **401**, 154–161.

Lacomba, R. *et al*. (2009) Determination of sialic acid and gangliosides in biological samples and dairy products: a review. *J. Pharm. Biomed. Anal.*, **51**, 346–357.

Massodi, M. *et al*. (2010) Comprehensive lipidomics analysis of bioactive lipids in complex regulatory networks. *Anal. Chem.*, **82**, 8176–8185.

McDonald, J.G. *et al*. (2007) Extraction and analysis of sterols in biological matrices by high performance liquid chromatography electrospray ionization mass spectrometry. *Methods Enzymol.*, **432**, 145–170.

Mukheree, K.D. and Weber, N. (eds) (1993) *Handbook of Chromatography, Analysis of Lipids*, CRC Press, Boca Raton, FL.

Murphy, R.C. *et al*. (2008) Imaging of lipid species by MALDI mass spectrometry. *J. Lipid Res.*, **50**, S317–322.

Sullards, M.C. *et al*. (2007) Structure-specific, quantitative methods for analysis of sphingolipids by liquid chromatography-tandem mass spectrometry: "inside-out" sphingolipidomics". *Methods Enzymol.*, **432**, 83–115.

Vaz, F.M. *et al*. (2015) Principles and practice of lipidomics. *J. Inherit. Metabol. Dis.*, **38**, 41–52.

Wenk, M.R. (2005) The emerging field of lipidomics. *Nat. Rev. Drug Discov.*, **4**, 594–610.

Zhao, Y.Y. *et al*. (2014) Lipidomics applications for discovering biomarkers of diseases in clinical chemistry. *Int. Rev. Cell Mol. Biol.*, **313**, 1–26.

25

Analysis of Post-translational Modifications: Phosphorylation and Acetylation of Proteins

Anja Stefanski,[1] Daniel Waldera-Lupa,[1] Gereon Poschmann,[1] Helmut E. Meyer,[2] and Kai Stühler[1]

[1]Heinrich Heine University Düsseldorf, Biologisch-Medizinisches Forschungszentrum (BMFZ), Molecular Proteomics Laboratory, Universitätsstraße 1, 40225 Düsseldorf, Germany
[2]Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V., Bunsen-Kirchhoff-Straße 11, 44139 Dortmund, Germany

During their lifetime, many proteins become co- or post-translational modified. On the one hand, they can become modified for a considerable time, for example, by the attachment of glyco-chains or lipids in specialized organelles like the endoplasmic reticulum or the Golgi apparatus. On the other hand, they are often chemically modified (phosphorylation, acetylation, methylation, etc.) in the context of signaling or metabolic processes. The occurrence of such post-translational modifications dramatically enhances the complexity of the proteome and it may appear that from a single protein coding gene sequence in addition to pre-translational mRNA splicing several hundred different protein species (modified variants) originate.

Biological processes like growth, development, differentiation, proliferation, and apoptosis of cells are mainly mediated by proteins and their interaction with each other and with metabolites, lipids, sugars, and nucleic acids. For example, proteins mediate the communication between cells by inter- and intracellular signal transmission, which is important for the coordinated behavior of different cell populations of an organism. Incoming signals were registered by specific receptor proteins sitting, for example, on the cellular surface, transferred to other proteins transducing the signal to more downstream oriented protein targets. Depending on the nature of the original signal, different types of signaling proteins become involved in signal transmission leading to well-defined signal cascades that control cells' behavior and fate. Thereby, the functions and activities of proteins are not only controlled by the available protein amounts, depending on the rates of gene expression, biosynthesis, and degradation, but are additionally modulated by specific post-translational modifications (PTMs). PTMs enhance the number of molecular structures enabled by the 20 directly coded proteinogenic amino acids building an additional level for controlling interaction, localization, activity, and stability of proteins. To date, more than 300 different types of PTMs have been identified. In the next section we will focus on phosphorylation and acetylation.

25.1 Functional Relevance of Phosphorylation and Acetylation

25.1.1 Phosphorylation

Reversible phosphorylation of distinct amino acids of proteins plays an important role in the regulation of intracellular signal cascades. Proteins will be phosphorylated by specifically acting kinases and dephosphorylated by specific phosphatases. Depending on the amino acid residue that is modified and the nature of the protein its activation or deactivation is either modulated by phosphorylation or de-phosphorylation or vice versa. There are several reasons why protein phosphorylation is well suited for the coordination of cellular answers as a reaction on a specific signal. On the one hand, the reaction is reversible and happens in only a few

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

seconds, facilitating a coordinated transfer of the signal into a biochemical answer. On the other hand, one protein can transfer the signal to a couple of downstream targets. Therefore, the activation of one single kinase might result in the phosphorylation of several proteins, leading to an efficient amplification of the original signal and a fast response.

In recent years, many details of the mechanism of phosphorylation have been discovered and a large number of protein kinases and kinase substrates identified. During phosphorylation, protein kinases build a complex with their protein substrates and catalyze the transfer of a phosphate group from a phosphate group donor to a specific amino acid residue of the substrate protein. Thereby, the γ-phosphate of ATP or an alternative nucleoside triphosphate or phosphoenolpyruvate acts as a donor of the phosphate group.

There are several mechanisms controlling the substrate specificity of protein kinases. Probably the most important one is the determination of substrate specificity by the amino acid sequence of the so called docking sites. These docking sites are specific, modular organized, and have a strong influence on the phosphorylation efficiency. The structural and function effect of a specific phosphorylation is mainly determined by its surrounding because the transfer of one additional phosphate group will introduce two additional negative charges into the substrate protein, eventually facilitating the formation of new salt bridges and hydrogen bonds. As a consequence, protein phosphorylation affects proteins' electrostatic interaction, binding properties, conformation, and catalytic activity and thereby influences the biological function of proteins. Four different types of phosphorylated amino acids have been described, varying in stability and function:

- *O-phosphate* (phosphorylation at hydroxyl residues of serine, threonine, and tyrosine);
- *N-phosphate* (phosphorylation at amino groups of arginine, lysine, and histidine);
- *S-phosphate* (phosphorylation at thiol groups of cysteines);
- *acyl phosphate* (phosphoanhydride form of aspartic acid and glutamic acid).

In biological systems, serine is the most commonly phosphorylated amino acid and the reversible O-phosphorylation of serine, threonine, and tyrosine residues represents one of the most important cellular regulation mechanisms in eukaryotic cells. Phosphorylation of serine and threonine mainly affects the regulation of enzymatic activity whereas tyrosine phosphorylation is important for signal transduction and the cellular answer to hormones, growth, and differentiation factors. N- and S-phosphates represent reactive intermediates of different biochemical reactions. For example, in *crustacean,* arginine phosphate plays an important role in the context of energy retrieval. In the bacterial phosphotransferase system, an energy-rich histidine phosphate is involved in the reaction of the HPr-kinase/phosphatase and a cysteine phosphate has been detected as an intermediate of the tyrosine phosphatase reaction. Acyl phosphate containing proteins play a role in the sensory transduction of the bacterial chemotaxis.

25.1.2 Acetylation

In contrast to well-established kinase mediated protein phosphorylation, the acetylation of proteins by acetyl transferases has emerged more and more as a biologically relevant modification during recent years. Acetyl transferases transfer acetyl groups from acetyl-coenzyme A to either the α-amino groups of amino terminal amino acids or the ϵ-amino group of internal lysine residues of proteins. Additional free thiol groups of cysteine residues have been described as becoming acetylated even in the absence of acetyl transferases. The irreversible amino terminal acetylation of proteins by N-acetyltransferases (NATs) is the most frequent PTM in eukaryotic organisms but is uncommon in prokaryotes. Its biological relevance depends mainly on the respective context. Whereas the reversible acetylation of internal lysine residues by histone acetyl transferases (HATs) has already been known for 40 years, the relevance of the internal lysine acetylation of non-histone protein emerged quite recently. The role of non-enzymatically mediated acetylation of cysteines is less clear but acetylated cysteines have been described as reactive enzyme intermediate *in vivo*.

Biochemically, the acetylation of a lysine residue within a protein not only neutralizes its charge but also compromises its ability to build hydrogen bonds and salt bridges. Accordingly, like phosphorylation, the acetylation of proteins affects their electrostatic interactions, binding properties, functions, stability, and subcellular localization. Because of the rising number of identified HAT-substrates, the discovery of protein acetylation reversibility by histone deacetylases

In vivo Latin for "within the living," describes processes in living organisms or living cells.

In vitro In contrast to *in vivo*, processes taking place in test tubes or more generally outside a living organism are characterized by the term *in vitro.*

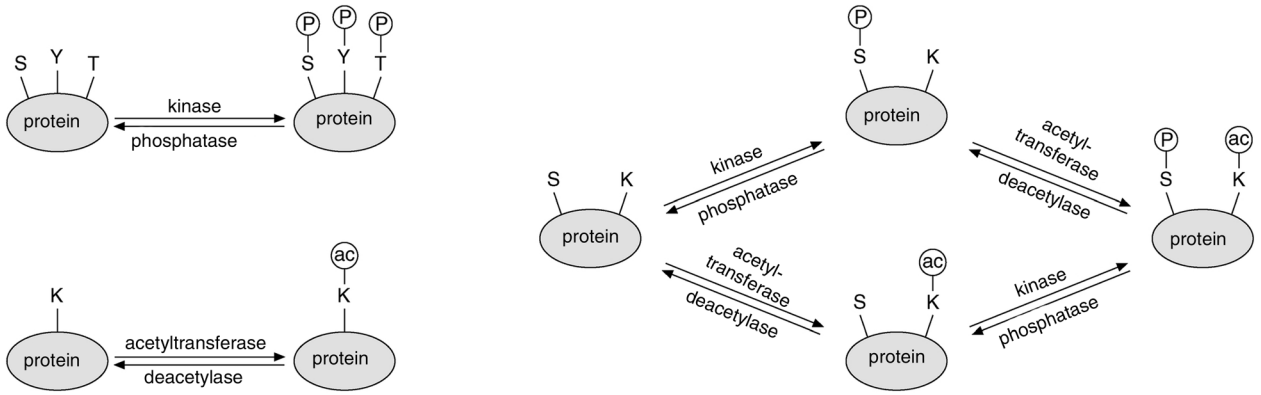


Figure 25.1 Post-translational modifications of proteins by different enzymes. Serine, threonine, or tyrosine residues in proteins can be phosphorylated by protein kinases and dephosphorylated by phosphatases. Lysine residues within proteins can be acetylated by (histone-) acetyltransferases and deacetylated by (histone-) deacetylases. Protein modification at multiple sites by several enzymes enables a dynamic control of cellular signal networks.

(HDACs), and the characterization of acetyl-lysine as molecular interaction signal, the detailed description of the acetylation status of internal lysine residues emerged as an increasingly important part of the puzzle describing the functional relationship of the different bioactive molecules within a cell. Meanwhile there are several hints that the above-mentioned principles of signal transmission by phosphorylation can also be applied to the acetylation of lysine residues because acetylation takes place within seconds, is reversible, and enables efficient signal amplification.

In recent studies on human cell lines about 3500 acetylated proteins on 1700 different proteins have been identified. Here not only proteins acting in the nucleus have been found but also cytosolic proteins acting predominantly in lager protein complexes. This underlies the importance of acetylation not only for nuclear localized but also for cytosolic functional aspects. Thereby, acetylation is important for a wide range of cellular functions including transcription, replication, cell cycles, chromatin dynamics, and apoptosis.

Several families of HAT enzymes and classes of HDAC enzymes are known that vary in their functional domains and catalytic mechanisms and modify specific lysine residues of their substrate proteins. As in transcription factors and transcription regulators a conserved protein module – the so-called bromodomain – has been identified that specifically recognizes and binds acetyl-lysines. The acetyl-lysine dependent protein–protein interaction resembles an important principle known from phosphorylation cascades: a modified residue generates a binding site for a protein module specifically recognizing this binding site. This principle was discovered about 20 years ago by investigating the interaction of phosphotyrosine and SH2 domains. The same principle can be applied to other modifications, for example, the binding of phosphoserine and phosphothreonine residues by 14-3-3, WW, or FHA-domains. As the interplay of several PTMs is a prerequisite of a dynamic and regulated signaling network, it has been suggested that bromodomain containing proteins recognize the molecular information of a lysine acetylation and pass it over to other proteins by interaction and modification mechanisms. In this model protein phosphorylation and acetylation represent switch-like signals from cellular signal cascades (Figure 25.1).

25.2 Strategies for the Analysis of Phosphorylated and Acetylated Proteins and Peptides

For the analysis of post-translational modifications there are still several challenges concerning sample handling and measurement. In principle, the methods applied routinely in proteome analytics for the identification and characterization of proteins can also be applied for PTM analysis. However, for several reasons the analysis is much more complex. Proteins are often

Proteome Analysis, Chapter 39

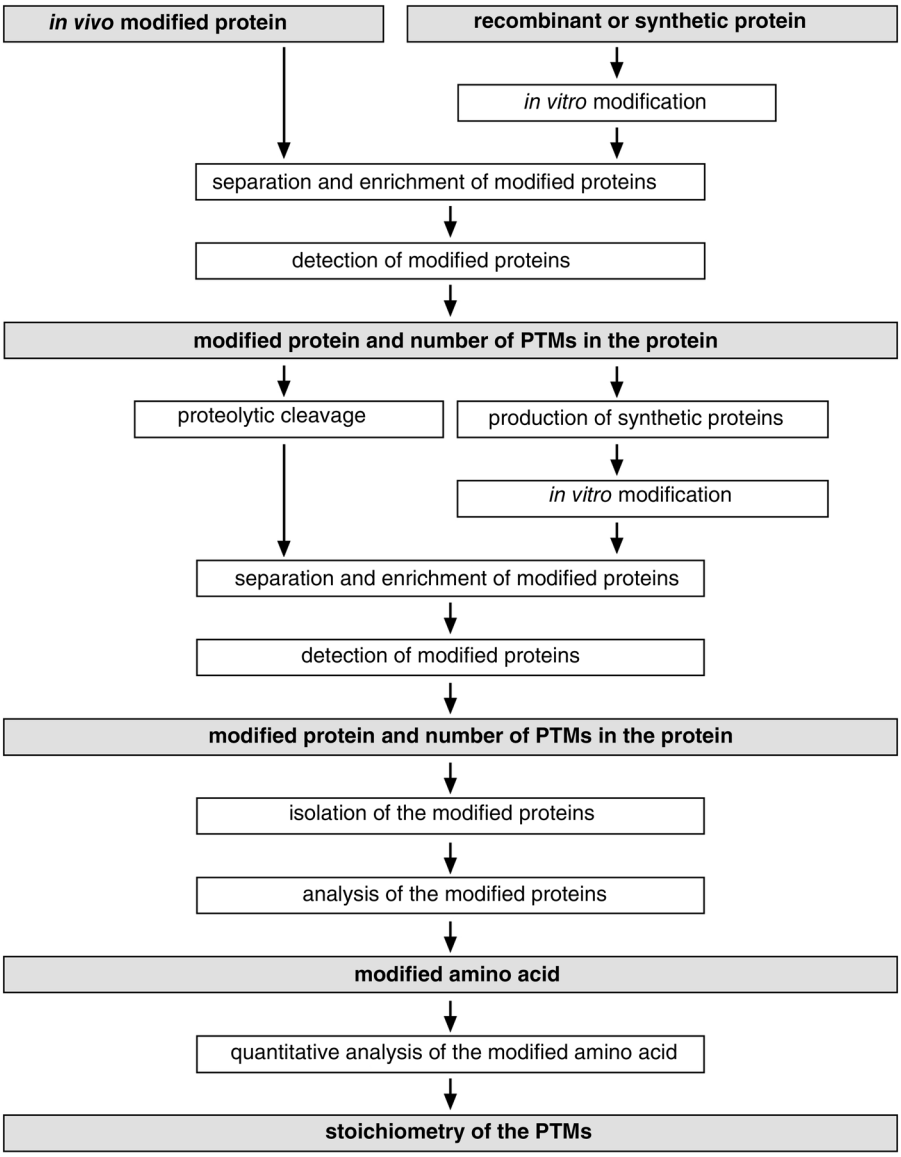


Figure 25.2 Analysis of post-translational modifications. There are four steps in the comprehensive analysis of *in vivo* or *in vitro* phosphorylated or acetylated proteins: the identity of the modified protein and the number of its PTMs are determined, the modified protein regions and the number of their PTMs are determined on the basis of peptides, the modified amino acids are identified, and the stoichiometry of the PTMs is quantified.

not completely modified, making a higher detection sensitivity or pre-analytical enrichment necessary. The determination of the exact site of modification within a protein or peptide can be hampered by the additional level of complexity as well as the different physicochemical properties like solubility, hydrophobicity, or molecular weight. Furthermore, it is important to pay attention to the chemical stability of a modified amino acid during the analysis, which might vary depending on the solvents used. The O-phosphates are the most stable protein phosphates and the amide bond of acetylated lysine residues is also known to be stable. This is the reason why mainly methods for the analysis of phosphoserine, -threonine, and -tyrosine as well as acetyl-lysine have been developed so far. Therefore, the analysis of these modifications will be the focus of the next section.

Methodically, the analysis of PTMs is based on the nature of the samples and the modification and can be structured in the following steps (Figure 25.2):

- modified proteins and peptides separated and/or enriched (Section 25.3);
- detection of post-translational modifications in proteins and peptides (Section 25.4);
- localization of modified amino acids (Section 25.5);
- quantitative determination of the stoichiometry of the PTMs (Section 25.6).

In the following sections, different strategies for protein analytic detection, analysis and localization of protein phosphorylation and acetylation will be explained in more detail.

25.3 Separation and Enrichment of Phosphorylated and Acetylated Proteins and Peptides

The analysis of post-translational phosphorylation and acetylation of proteins is carried out either on *in vivo* modified proteins or on proteins modified *in vitro* by respective kinases or acetyl transferases. Generally the subsequent successful detection of post-translational modifications depends largely on the stoichiometry of the PTMs. In a cell a reversible *in vivo* PTM normally results in sub-stoichiometric modifications – only one or two percent of a protein exist in their phosphorylated or acetylated form. If the stoichiometry lies below the sensitivity limit of the detection method, a *separation, enrichment or purification* of the modified protein is mandatory. This enables a detailed characterization to be made independently of unmodified proteins and interfering compounds. Methods often used for protein separation and fractionation are liquid chromatography (LC) and one- or two-dimensional polyacrylamide gel electrophoresis (1D, 2D PAGE). In recent years, mass spectrometry has emerged more and more as an important method for the qualitative and quantitative analysis of protein modifications. For the mass spectrometric based analysis of a protein modification it is favorable to cleave the protein into peptides using a suitable endoprotease. As limiting factors the low amount of modified analyte compared to its unmodified counterpart as well as the enhanced complexity still remain.

Chromatographic Separation Techniques, Chapter 10

Electrophoretic Techniques, Chapter 11

For the specific isolation and enrichment of phosphorylated and acetylated proteins and peptides, a couple of different methods exist that make use of the different mobility, affinity, or binding properties of the modified protein or peptide. Alternative approaches fuse a tag to modified analytes, facilitating enrichment or detection. An important method for the purification of phosphorylated and acetylated proteins is immunoprecipitation (IP). Particularly for the analysis of phosphorylated proteins and peptides affinity chromatography based methods with metal chelates and metal oxides as well as strong cation exchange (SCX) chromatography work well. Furthermore, highly specific enrichment methods have been developed in the context of modification specific analytics like two-dimensional phosphopeptide mapping (2DPP-mapping) and β-elimination with subsequent Michael addition. Moreover, classical chromatographic separation methods like nano reversed phase (RP) HPLC and capillary electrophoresis (CE) are applied to achieve an efficient separation of complex peptide mixtures.

Immune Precipitation, Section 5.3.2

Capillary Electrophoresis, Chapter 12

Immunoprecipitation makes use of modification specific antibodies to detect phosphorylated and acetylated proteins. Normally, the specific antibody has been bound to an immobilized matrix. After complexes of modified proteins and antibodies have been formed and with the accompanying precipitation of the protein, a stringent washing and removal of unmodified proteins is feasible. Subsequently, the immunoprecipitated protein is separated from the antibody and potentially co-precipitated but unmodified proteins by LC or 1D or 2D PAGE to lower sample complexity for the subsequent PTM detection and analysis. The successful enrichment of post-translational modified proteins strongly depends on the quality of the antibody used and the strength of the protein–antibody affinity. Using modification specific antibodies, immunoprecipitation can be successfully applied for the enrichment of different species of O-, N-, S-, and acyl phosphate as well as for acetylated proteins.

In contrast, immobilized metal chelate chromatography (IMAC) is useful for the enrichment of phosphorylated proteins and peptides independent of the nature of the phosphorylated amino acid. IMAC makes use of the interaction between negatively charged phosphate groups and positively charged immobilized metal ions (chelated Fe(III) or Ga(III) ions). A disadvantage of the method is the unspecific binding of acidic proteins at the immobilized metal ions, which can be extenuated by methyl esterification of carbonyl groups of proteins respective peptides. Beside IMAC, chromatography with metal oxides like titanium or zirconium oxide has been established as an efficient method for the enrichment of phosphopeptides. The principle is also based on an affinity-chromatographic enrichment whereas TiO_2 builds up a bidentate surface complex with phosphate groups. Phosphoanalytic methods make use of TiO_2 coated magnetic particles as well as chromatographic nanoparticles that exhibit an enormous surface area.

Another strategy for selective enrichment of phosphopeptides is strong cation exchange (SCX)-chromatography. It is based on the competitive interaction of the positively charged analyte (peptides) with the negatively charged stationary phase under acidic conditions. The displacement and elution of peptides is accomplished by an increasing salt concentration of the eluent. The higher the positive charge of the peptide, the more strongly it binds. Therefore, phosphopeptides elute due to the additional negative charge of the strong acidic phosphate group already at low salt concentrations from the ion exchanger and thus can be selectively enriched. SCX chromatography can also be used for the purification of phosphorylated proteins.

All the methods for the specific enrichment or purification of modified proteins or peptides described above require multiple processing and desalting steps. A combination of these methods with the temporary immobilization of the analyte on a suitable support material makes sense, because multiple reactions and washing steps can be performed with maximum efficiency but minimal sample loss. Suitable carrier materials are mainly solid phases and particles with porous surfaces, functionalized with 3, 8, or 18 carbon units (C_3, C_8, C_{18}) long hydrophobic alkyl moieties (reversed phase, RP).

The 2DPP mapping, a two-dimensional method for the isolation of phosphopeptides, is more of historic relevance. Phosphopeptides are separated, regardless of the nature of the phosphorylated amino acid, by thin-layer chromatography in the first dimension according to their electrophoretic mobility and in the second dimension according to their hydrophobicity. After visualization of the separated peptides by radioactive or fluorescence-based methods, each spot represents a phosphorylated peptide similar to the bead chains patterns formed in 2D-PAGE. The applicability of this technique for the separation of acetylated peptides has not yet been shown.

A further interesting method for the enrichment of phosphopeptides with a more specific application is β-elimination followed by Michael addition. By combining β-elimination and Michael addition, the position of the modification is fixed by the covalent attachment of a linker group. The introduction of an affinity group such as a biotin tag offers, additionally, the possibility of peptide enrichment by affinity chromatography (Figure 25.3). During the β-elimination the base-catalyzed removal of the phosphate group from the phosphorylated amino acid residue takes place, resulting in an elimination of phosphoric acid (H_3PO_4). In the second step, the Michael addition, the amino acid residue is derivatized with a thiol-containing substances (HS-X). These substances may also contain, in addition to biotin, a combination of affinity and isotope labeled groups such as ICAT (isotope-coded affinity tag) for subsequent relative quantitative analysis.

The Michael addition can be facilitated with monothiols or with bifunctional thiols. The signal shift of a peptide caused by the addition after derivatization can be easily detected by means of MS and interpreted as evidence for the existence of a phosphorylation site. Regardless of the labeling group, however, this method is suitable only for serine or threonine phosphorylated peptides because only they allow a β-elimination. In addition, due to the very basic conditions of the β-elimination, partial racemization of peptides takes place, which makes a subsequent chromatographic separation of a complex peptide mixture, as usually done in proteome analysis, almost impossible.

A comprehensive purification method for acetylated proteins and peptides does not exist yet. In addition, methods based on the conversion of the chemically modified protein into a more easily cleaned or detected species focus on phosphorylated proteins. Generally, the techniques to separate complex protein or peptide compositions used in proteome analysis can be used here, as well as for phosphorylated proteins/peptides. This includes chromatographic separation methods such as nano-RP (reversed phase) HPLC and capillary electrophoresis (CE) for the separation of peptides, or one - or two-dimensional polyacrylamide gel electrophoresis (1D-, 2D-PAGE) for protein separation. Nano-RP-HPLC has a particularly broad application for separation of the modified peptides. Nano-HPLC offers the possibility to separate very low concentrations of peptide as well as the opportunity of online coupling with an ESI mass spectrometer. This prevents peptide losses through repeated manual handling of samples and is easy to adjust in an automated manner to the respective PTM analysis. Capillary electrophoresis (CE) is carried out carrier-free in a capillary and uses aqueous buffer systems as separation media. Due to the small sample volumes, the good separation efficiency, the high resolution of peptide signals, and the possibility of direct coupling to the MS, CE provides comparable advantages for the isolation of phosphorylated and acetylated peptides as the nano-HPLC.

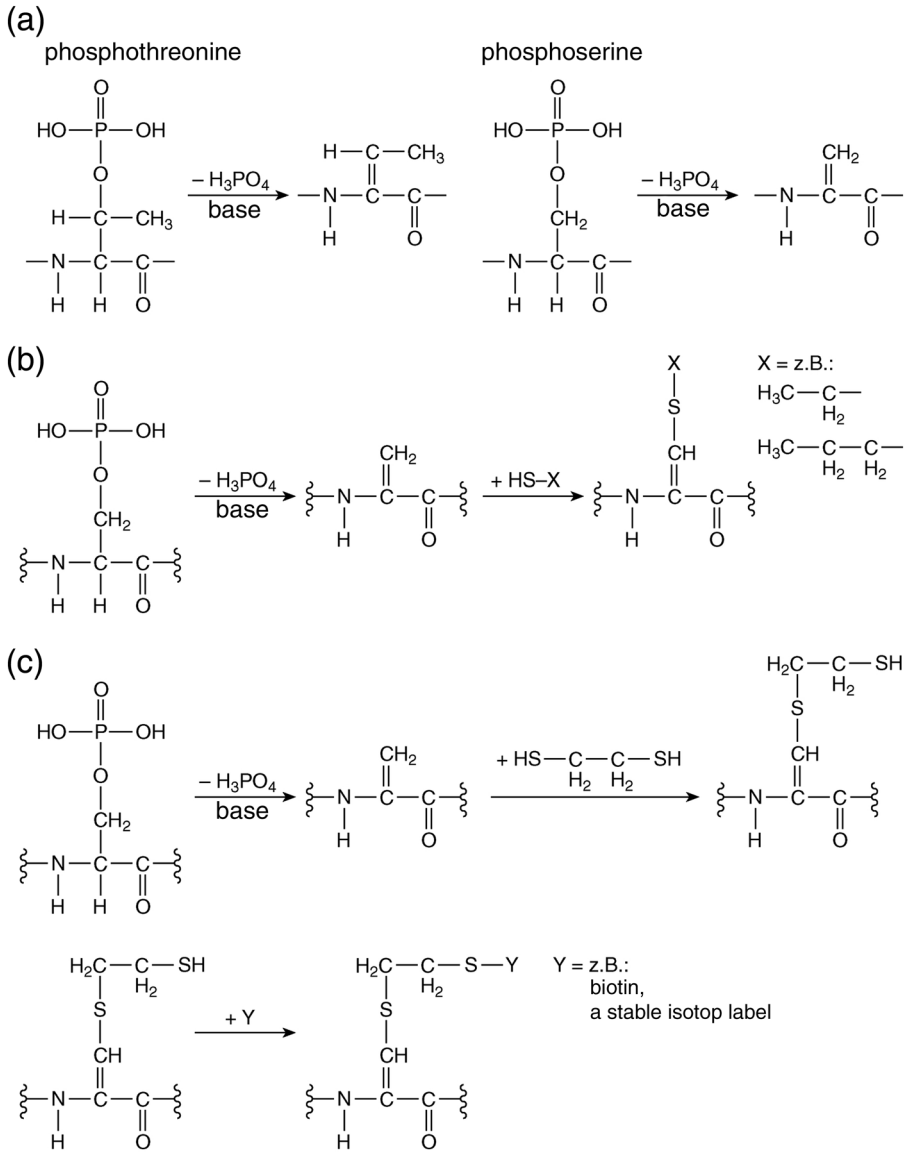


Figure 25.3 Base-catalyzed β-elimination and Michael addition with phosphoserine and phosphothreonine. (a) By treatment with a strong base, the phosphate group of phosphoserine and phosphothreonine can be eliminated. (b) The resultant dehydroalanine or dehydro-L-amino-2-butanoic acid group reacts with thiol compounds such as ethanethiol or propanethiol simply according to the mechanism of the Michael addition. (c) Additionally, these thiol containing substances may have affinity tag groups such as biotin, a stable isotope label, or a combination of both groups (e.g., ICAT).

25.4 Detection of Phosphorylated and Acetylated Proteins and Peptides

25.4.1 Detection by Enzymatic, Radioactive, Immunochemical, and Fluorescence Based Methods

The specific detection of modified proteins can be accomplished by treatment with phosphatase or deacetylase and using radioactive, immunochemical or fluorescence-based methods. During detection in the 2D gel, modified proteins can be detected as a characteristic spot pattern, since any phosphorylation and acetylation of a protein lowers its isoelectric point and increases its molecular weight, which can mostly be seen by a change in migration behavior of the proteins in PAGE (Figure 25.4).

These shifts in protein migration are indicators of multiple modified proteins, but can also occur by the introduction of differences in charge by deamidation of asparagine or glutamine during sample preparation.

Western Blot, Section 5.3.3

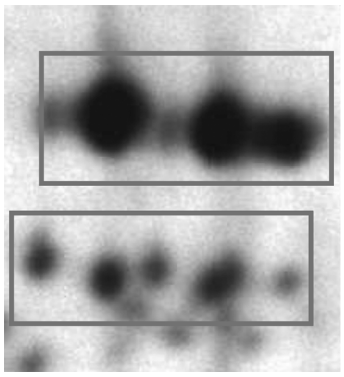


Figure 25.4 Characteristic spot patterns resulting from different phosphorylation status of proteins. By the introduction of a highly negatively charged phosphate group the isoelectric point of the protein is shifted into the acidic region and prevents the binding of SDS by repulsive Coulomb interactions, so that the apparent molecular weight increases. The result is a characteristic spot pattern, with the introduction of an additional phosphate group leading to a shift of the protein pH by about 0.3 units.

To check the reason for the separation of a protein into a variety of spots, the sample can be treated prior to 2D-PAGE with phosphatases or deacetylases, since complete dephosphorylation/deacetylation of a protein reverses the changes to its isoelectric point and its molecular weight made by the PTM and therefore the protein should show up only in a single spot. A parallel 2D-PAGE of untreated and phosphatase/deacetylase-treated sample allows a differential analysis of the resulting gel images for the presence of phosphorylated or acylated proteins.

Radioactive detection methods require the PTM of the protein to be carried out using radioactive marker groups. They are limited to samples from *in vivo* labeling studies or from *in vitro* modification of recombinant or synthetic proteins. Marker group donors are usually ^{32}[P]-ATP or ^{33}[P]-ATP for phosphorylation and ^{14}C acetyl coenzyme A for acetylation. After separation of the radioactively labeled proteins by LC or PAGE, modified proteins are detected by scintillation, Cherenkov counting, or autoradiography (Figure 25.5). With autoradiography, all phosphorylated or acetylated proteins are detected, but no information is provided about the nature of the modified amino acid, which means one cannot distinguish between O-, N-, S-, or acyl phosphates or between acetyl-lysine and unspecific acetylated amino acids.

Alternatively, modification-induced generation of protein isoforms into a number of spots can be detected by Western blot analysis using modification-specific antibodies. During Western blot analysis, proteins, separated by a polyacrylamide gel, are transferred to a membrane made of poly(vinylidene fluoride) (PVDF) or nitrocellulose. After blocking nonspecific binding sites the membrane is incubated with the modification-specific primary antibody. Information about which PTM is present in the modified protein is obtained by the choice of the primary antibody. Subsequent visualization with a secondary antibody directed against the first antibody is carried out by a color reaction, chemiluminescence, or fluorescence. Immunochemical detection of PTMs requires no special treatment of the samples, however, as the IP strongly depends on the specificity and binding affinity of the applied antibodies. An increasing number of modification and sequence specific antibodies for the analysis of phosphorylated and acetylated proteins are commercially available and can be used for detection of the modified proteins. Especially, phosphotyrosine antibodies are specific enough to detect a single phosphorylated tyrosine independently of the adjacent amino acids, thus allowing the generic detection of phosphotyrosine-containing proteins.

Immunological Techniques, Chapter 5

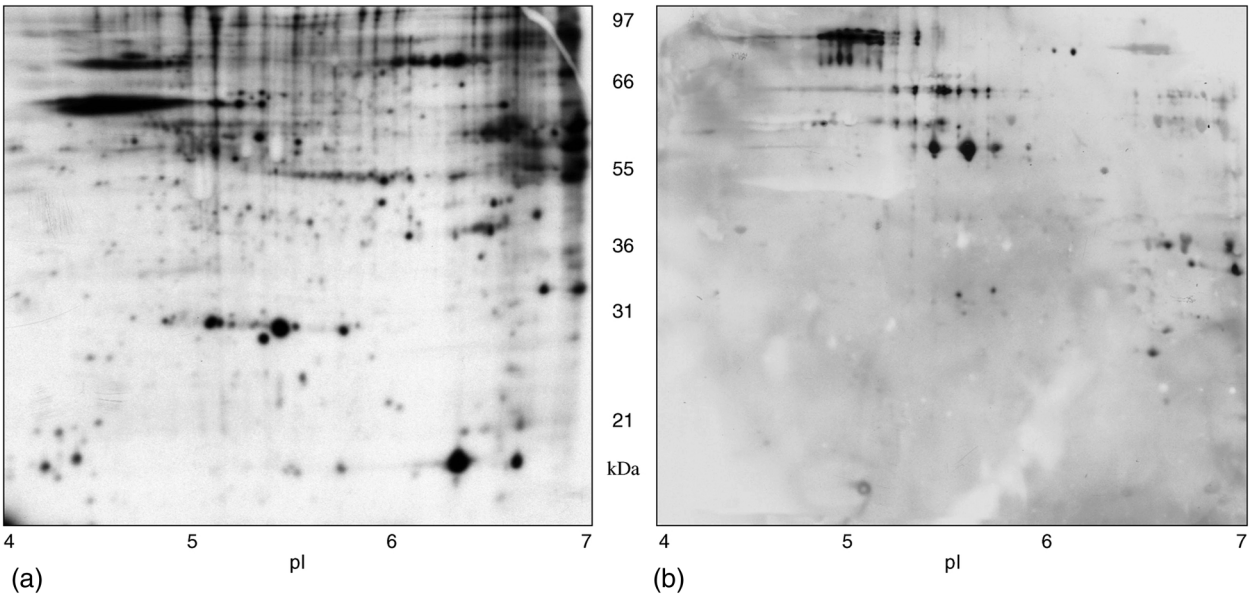


Figure 25.5 Autoradiograph-based Western blot analysis after 2D-PAGE. (a) Autoradiogram of a 2D gel of human platelet proteins in the pH range 4–7 after metabolic labeling with ^{32}[P]. On the basis of the autoradiogram no information is provided about the type of phosphorylation (N-, O-, S-, or acyl-phosphate). For a subsequent precise specification detection with specific antibodies is suitable. (b) Western blot analysis using an anti-phosphotyrosine antibody of the autoradiogram shown in (a).

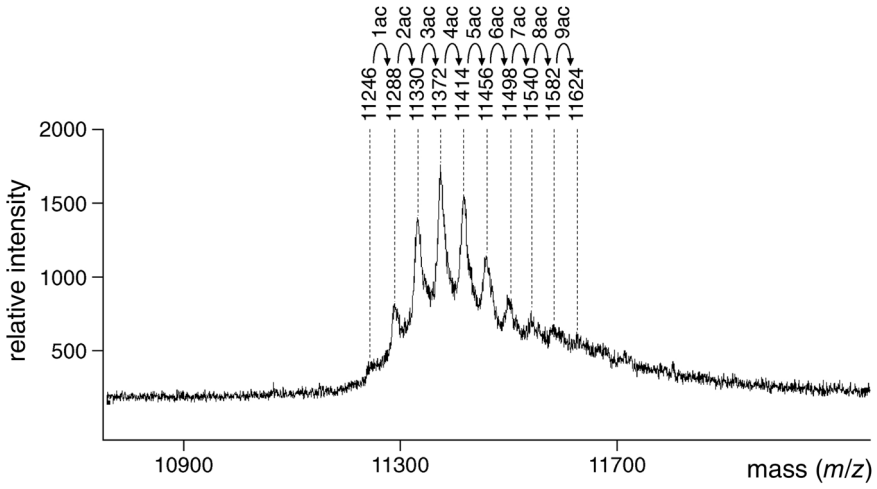


Figure 25.6 MALDI-TOF mass spectrum of chemically acetylated histone H4 (recombinant *Xenopus laevis* H4, courtesy of K. Tóth and J. Langowski, DKFZ Heidelberg). The shift of the protein mass (*m/z* 11246) by nine steps of +42 Da each indicates the acetylation (AC) of nine lysines. Because the protein sequence does not contain any cysteine a non-enzymatic acetylation of cysteines can be excluded.

Fluorescence-based detection methods have been developed for phosphorylated proteins, but not so far for acetylated proteins. They use the non-covalent binding of small organic fluorophores to phosphorylated amino acid residues to enable direct and sequence-independent detection of phosphoproteins in polyacrylamide gels. Detection is possible without any special pretreatment of the samples, but requires a specific fluorescence scanner. Their sensitivity is dependent upon the binding of the fluorophores to the phosphorylated amino acids and the stoichiometry of phosphorylation.

Mass Spectrometry, Chapter 15

25.4.2 Detection of Phosphorylated and Acetylated Proteins by Mass Spectrometry

The radioactive, immunochemical, or fluorescent based detection of post-translationally modified protein commonly provides only information on the presence of PTMs, while mass spectrometric detection methods using MALDI or ESI MS give also information on the number of PTMs in a protein. The PTM of a protein leads to a characteristic increase in its molecular weight: the increase upon phosphorylation is +80 Da and upon acetylation it is +42 Da. Therefore *in vivo* or *in vitro* modification of a protein can be detected by determination of its exact mass using MS and identified by comparison of the measured weight with the theoretical mass of the unmodified protein. The difference between measured and theoretical mass of a protein can also reveal if a protein has multiple PTMs of the same or different type (Figure 25.6). Particularly important is the determination of the amount of PTMs during the analysis of acetylated proteins, because a non-enzymatic and sequence-independent acetylation of all cysteine residues of the protein takes place by acetylation by AcCoA. As the acceptor acetyl group of cysteines is often not considered or is mistaken for autoacetylation of proteins, acetyltransferases for a series of substrates whose actual enzymatic acetylation must be regarded as unlikely can be found in the literature. The non-enzymatic acetylation of cysteines should therefore always be considered when analyzing enzymatic lysine acetylation. A distinction between the two acetylation types by mass spectrometry is possible by comparing the number of bonded acetyl groups with the number of contained lysine and cysteine residues. Both forms can also be distinguished by sequencing of the protein or by Western blot analysis using an acetyl-lysine specific antibody.

25.5 Localization and Identification of Post-translationally Modified Amino Acids

To localize a modified amino acid in a peptide complete sequencing using Edman degradation or mass spectrometric analysis of fragment ions can be applied. Sequencing of peptides is

particularly susceptible to interference by other peptide species and therefore requires specific enrichment and isolation of such phosphorylated or acetylated peptides using the methods described above (Section 25.3).

25.5.1 Localization of Phosphorylated and Acetylated Amino Acids by Edman Degradation

Protein Sequence Analysis, Chapter 14

In classical Edman degradation, amino-terminal residues are sequentially labeled with phenyl isothiocyanate (PITC) and cleaved from the peptide without disrupting the peptide bonds between other amino acid residues. The generated stable phenylthiohydantoin (PTH)–amino acid derivatives are separated by a RP-HPLC column and identified according to their specific retention time. Edman degradation is the simplest method for the localization of phosphorylated or acetylated amino acids. However, successful application depends on the homogeneity of the peptide sample, the stoichiometry of the PTM, and the stability of the modified amino acid during derivatization. Inhomogeneous samples lead to simultaneous detection of several peptide sequences and impede the unambiguous localization of a modified amino acid in one of the peptides. In addition, the low stoichiometry of *in vivo* PTMs in combination with the decreasing repetitive yield of sequencing leading to the reduction in the concentration of the PTH derivative below the detection limit complicates the analysis. Moreover, modified amino acids and their derivatives have different chemical stabilities, therefore a specific adaptation of the sequencing protocol according to the PTM is necessary.

Due to their hydrophobicity phosphotyrosines can only be analyzed by solid-phase sequencing, and their PTH derivatives exhibit poor absorption and a variable retention. Phosphoserines and -threonines form dehydroalanine and dehydro-α-amino-2-butanoic acid by acid catalyzed β-elimination during Edman degradation, and can only be indirectly detected via dithiothreitol adducts of their respective PTH derivatives. An additional way to determine the localization of phosphoserines and -threonines using Edman degradation is by conversion into their stable derivatives 5-ethylcysteine and β-methyl-S-ethylcysteine by β-elimination and addition of mercaptoethane, which can be identified according to the specific retention time.

By using Edman degradation the localization of acetyl-lysines can be determined directly and distinguished from unmodified lysines. The PTH derivate of an unmodified lysine exhibits the highest hydrophobicity compared to the PTH derivatives of the 20 proteinogenic amino acids, because it carries two chromophores, namely, a PTH group by the modification of the amino terminus and a PTC group by modification of the ε-amino group of lysines' side chain. For this reason, it exhibits the longest retention time on a RP-HPLC column. In contrast, the PTH acetyl-lysine has a shorter retention time, because the ε-amino group is protected from PTC modification by acetylation, whereby the PTH derivate carries only one chromophore (Figure 25.7). The determination of unmodified or acetylated cysteine residues by Edman degradation is not possible. Unmodified cysteines react with the reagent PITC and form *N*,*S*-di-(phenylthiocarbamoyl) cysteine, which cyclizes under acidic conditions to 1-phenyl-2-thiono-4-(*N*-phenyldithiocarbamoylmethyl)-5-oxadiazole and is no longer detectable with RP-HPLC. This cyclization does not occur with acetylated cysteines. However, during Edman degradation acetylated cysteines undergo an intramolecular transfer of the acetyl group on the free terminal amino group, which is formed during cleavage of the previous amino acid. Thus, the N-terminus of the peptide is blocked for further reaction with PITC, whereby sequencing stops.

25.5.2 Localization of Phosphorylated and Acetylated Amino Acids by Tandem Mass Spectrometry

Mass Spectrometry, Chapter 15

For mass spectrometric based localization of PTMs it is usually necessary to perform a proteolytic cleavage of the proteins prior to the analysis. Thus, the average length and type of the resulting peptides can be influenced by selecting the endoproteinase. Commonly, the endoproteinase trypsin is used for the analysis of PTMs, whereby peptide bonds are cleaved at the carboxyl side of lysines and arginines. Due to the average distribution of lysines and arginines in proteins, mainly peptides with a length of 5–20 amino acids are generated carrying a positively charged residue at the carboxy terminus (K or R). The phosphorylated and

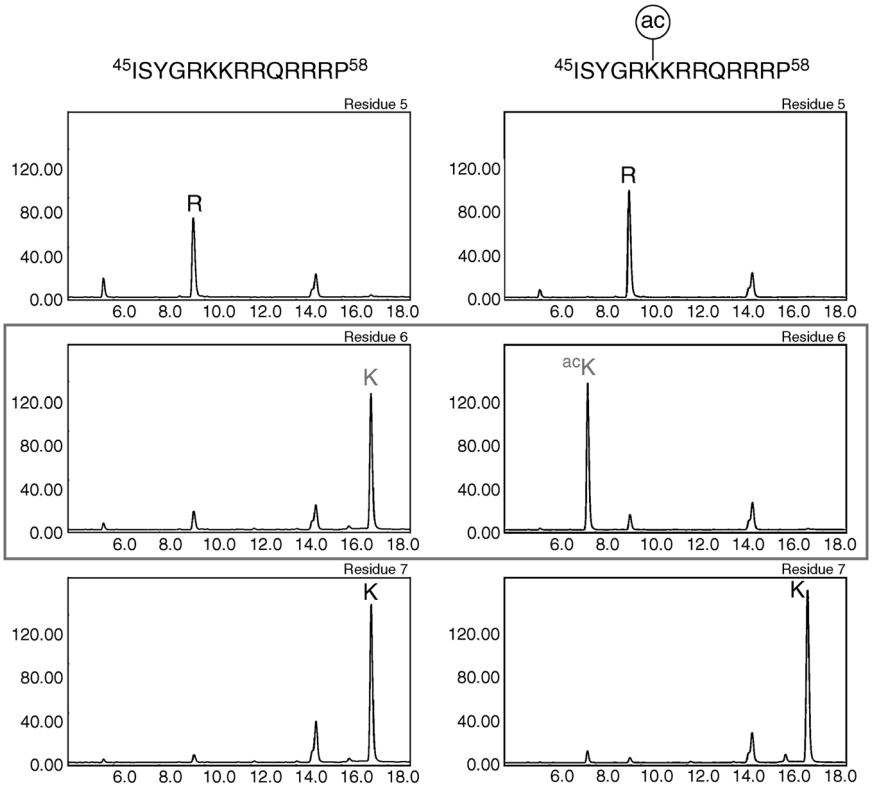


Figure 25.7 Identification of acetylation sites in proteins by Edman degradation. The retention time of lysines is significantly reduced by acetylation. The figure shows the degradation steps 5–7 of the unmodified and the acetylated (K50) peptide ^{45}ISYGRKKRRQRRRP^{58} of the HIV-1 Tat protein.

acetylated peptides are usually characterized by MALDI-MS or ESI-MS. These techniques provide information on the number of PTMs in a peptide and allow for a subsequent fragmentation and sequencing of the modified peptides. As a consequence, identification and localization of modified amino acids can be carried out without any further treatment of the sample. Technical improvements to the mass spectrometers, the automatization of mass spectrometric analyzes, and the various methods for the separation and enrichment of modified proteins and peptides have revolutionized the analysis of PTMs. The analysis of phosphorylated or acetylated peptides using mass spectrometry is based on two principles: the detection of an increase in molecular weight, which is characteristic for a PTM (Section 25.4.2), and the detection of the fragmentation pattern of the peptide, which allows the exact localization of the PTM. Due to the additional negative charge of the phosphate group and the parameter settings (positive mode) of the mass spectrometer, phosphorylated peptides exhibit a lower degree of ionization in the presence of unmodified peptides, which can lead to suppression or even loss of the signals. In addition, phosphorylated peptides are often unstable in the mass spectrometer and show a characteristic dissociation of phosphate in the form of H_3PO_4 and HPO_3, which can be used for identification. The loss of H_3PO_4 leads to a mass shift of −98 Da in singly and −49 Da in doubly charged ions. Especially, phosphoserines and -threonines show intense signals for the dissociation of H_3PO_4, which correspond to the formation of dehydroalanine from phosphoserine and dehydro-α-amino-2-butanoic acid from phosphothreonine. The dissociation of HPO_3, which leads to a mass shift of −80 Da in singly and −40 Da in doubly charged ions, is less frequently observed in phosphoserines and -threonines. Due to the conjugated system of the benzene ring, phosphotyrosines are stable phosphates and rarely exhibit the dissociation of HPO_3, but do exhibit the dissociation of H_3PO_4.

In contrast to phosphorylated peptides, amino terminally or internally (lysine) acetylated peptides are stable during MS analysis. In addition, they are more easily ionized and can be detected in the presence of unmodified peptides. Acetylated lysines exhibit neither in MALDI-MS nor in ESI-MS characteristic dissociations, because of the amide bond between the lysine residue and the acetyl group. Acetylated cysteines, where the acetyl group is bound as a thioester, are also stable during ionization using MALDI-MS or ESI-MS (Figure 25.8). For analysis of acetylated proteins, the specificity of the endoproteinases trypsin and Lys-C can be

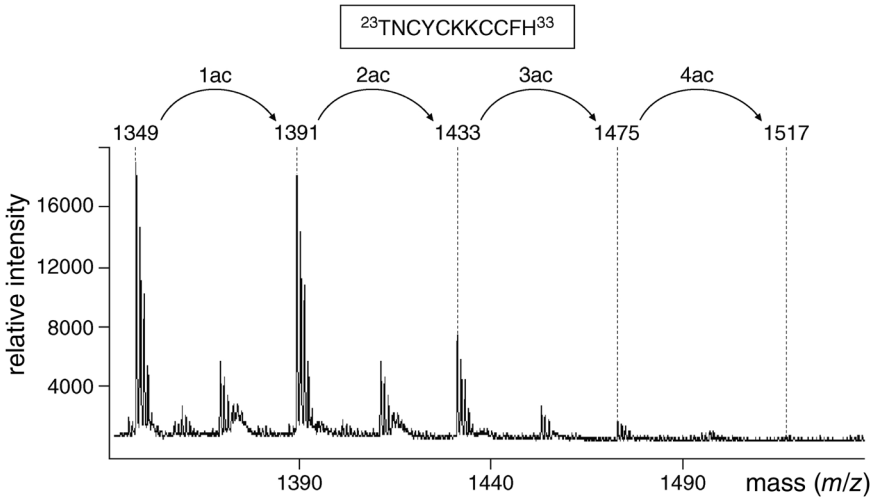


Figure 25.8 Cysteines are highly reactive acetyl group acceptors. MALDI-TOF mass spectrum of the peptide ^{23}TNCYCKKCCFH^{33} of the HIV-1 Tat protein after incubation with acetyl-coenzyme A. The four-step shift of the peptide signal (*m/z* 1349), each step by +42 Da, indicates acetylation (ac) of at least four amino acid residues, although the peptide contains only two lysines and the incubation was carried out in the absence of acetyltransferase. Thus, the four-step shift is based solely on the non-enzymatic acetylation of the cysteines. Source: Dormeyer, W. et al. (2003). *Anal. Bioanal. Chem.*, 376, 994–1005. With permission, Copyright © 2003, Springer-Verlag.

exploited. Only peptide bonds on the carboxy terminal side of unmodified lysines are cleaved by trypsin and Lys-C, whereas acetylated lysines are resistant to cleavage. Consequently, the acetylation site can be determined based on the specific fragmentation pattern and the position of the blocked cleavage site (Figure 25.9).

Due to the smooth ionization in ESI-MS, the loss of H_3PO_4 and HPO_3 is only caused by increasing the applied voltage or by induced collision with gas molecules, which also results in

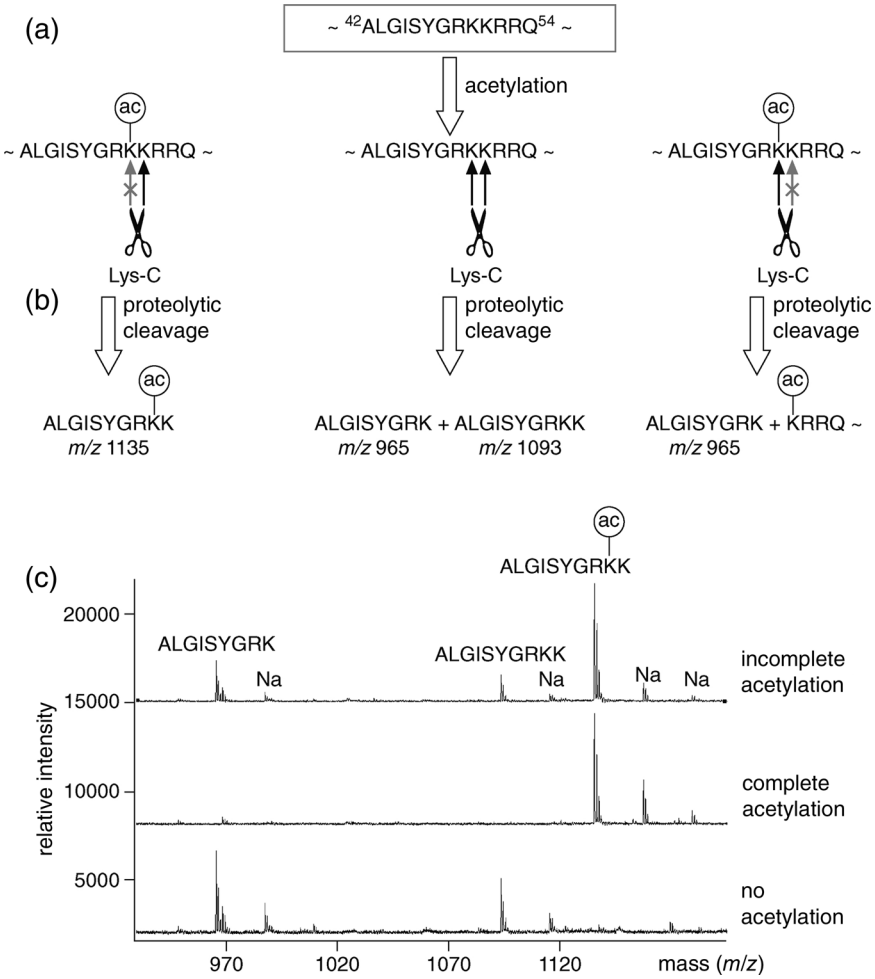


Figure 25.9 Identification of acetylation sites in proteins by cleavage with Lys-C. The cleavage pattern of a protein is affected by the acetylation of lysines, as Lys-C only cleaves peptide bonds on the carboxyl side of unmodified lysines. The figure shows the schematic cleavage using Lys-C and the detected fragmentation pattern of the peptide ^{42}ALGISYGRKKRRQ^{54} of the HIV-1 Tat protein using MALDI-TOF-MS. The protein is acetylated at lysine 50, but not at lysine 51. The signals of cleaved peptides as well as sodium adducts (Na^+) are marked. Source: courtesy of Martina Schnölzer, DKFZ, Heidelberg, Germany.

complete fragmentation of the peptide. Thus, localization of phosphorylated amino acids requires a fragment ion analysis. Using MS-based fragment ion analysis, the specific fragmentation pattern of a peptide is detected. Relying on the fragment ion spectrum, also known as the MS/MS spectrum, the presence and localization of a PTM can be determined. Therefore, different fragmentation techniques can be applied. Beside collision-induced dissociation (CID) alternative techniques, such as electron-induced fragmentation, are available for analysis of PTMs. For this purpose all mass spectrometers suitable for fragment ion analysis, such as TOF, ion traps, triple quadrupoles, and hybrids, can be applied for localization of PTMs. Among these, triple quadrupole MS with different modification-specific MS/MS analysis techniques plays a long standing role in PTM analysis. These experimental techniques are known as product ion scan, neutral loss scan, precursor ion scan, and PTM specific immonium ion scan.

Mass Spectrometry, Chapter 15

The product ion scan is the most commonly used MS/MS technique in triple quadrupole MS analysis (Figure 25.10a). Here, the parameters of the first quadrupole (Q1) are set to specific values such that only ions of a specific mass can pass into the second quadrupole (Q2). The other peptide ions are deflected and do not reach Q2. In Q2, fragmentation by low-energy collisions with a collision gas (nitrogen, helium, or argon) occurs (CID) breaking the peptide bonds, preferably breaks generating b- and y-ions, which are subsequently analyzed in the third quadrupole (Q3). The product ion scan produces a peptide specific fragment ion spectrum, which can be interpreted manually or by the use of search algorithms, thereby identifying the sequence of the peptide. Phosphorylated or acetylated amino acids can be identified due to the mass shift of +80 Da or +42 Da of the fragment ion signals in the MS/MS spectrum. In addition, phosphopeptides exhibit the above-mentioned dissociation of H_3PO_4 and HPO_3 in the MS/MS spectra starting at the phosphorylated residue. In contrast, acetylated peptides are stable and produce a characteristic mass shift of +42 Da, but no specific product ion signals.

In the precursor ion scan, Q1 scans across a mass range to determine the masses of all peptide ions (Figure 25.11). Subsequently, the peptide ions are sequentially transferred into Q2 and fragmented by CID. In Q3 the resulting fragment ions are detected. Here, only modification-specific fragment ions are of interest such as 79 Da for a phosphate group (PO_3^-). This specific fragment can be assigned to a precursor ion and provides information about the state of modification of the peptide. Using the precursor ion scan, the peptide ion carrying a phosphate group is initially determined. To determine the sequence of the peptide and thus the phosphorylated amino acid a fragment ion spectrum of the peptide is generated using the product ion scan. Due to their stability, acetylated and tyrosine-phosphorylated peptides exhibit no detectable specific fragment ions by precursor ion scan.

The neutral loss scan is an effective method for the identification of modified peptides. Here, the loss of a neutral group from a modified peptide is detected after low-energy collisions (Figure 25.10c). First, the masses of ionized peptide ions are determined in Q1; the ions are then

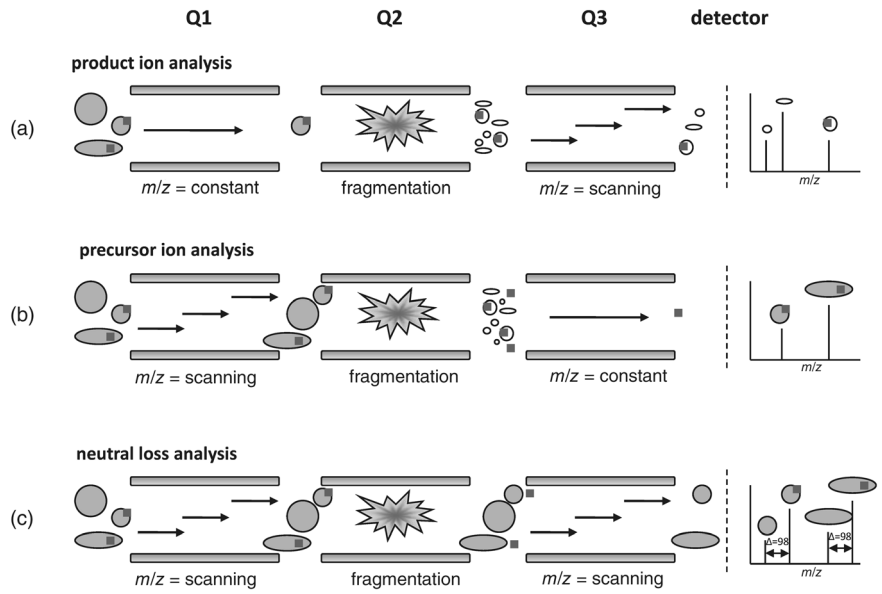


Figure 25.10 Fragment ion analysis. (a) Product ion scan: The electric field of Q1 is set to specific parameters such that only precursor ions of a specific *m/z* ratio can pass into Q2. Thereby, precursor ions can be selected for fragmentation into product ions, which are subsequently analyzed in Q3. This results in a fragment ion spectrum, where characteristic dissociations of H_3PO_4 (singly charged: −98 Da, doubly charged: −49 Da) and HPO_3 (singly charged: −80 Da, doubly charged: −40 Da) can be obtained. (b) Precursor ion scan: Initially, the *m/z* ratios of the peptide ions are determined in Q1. Afterwards, they are transferred into Q2, where they are fragmented into product ions using CID. In the precursor ion scan, Q3 is adjusted so that only PO_3^- ions with an *m/z* ratio of 79 Da are detected. (c) Neutral loss scan: Here the specific loss of a neutral fragment can be detected. In Q1, the *m/z* ratios of the peptide ions are determined, and the ions are transferred into Q2, where they are fragmented and the generated fragment ions are passed into Q3. During phosphopeptide analysis, Q3 scans for specific masses that are −98 Da (singly charged), −49 Da (doubly charged), and −32.7 Da (triply charged) lower than the original masses scanned in Q1. Using neutral loss scan, only serine and threonine phosphates can be detected. Due to their stability, acetylated and tyrosine-phosphorylated peptides exhibit no detectable specific fragment ions.

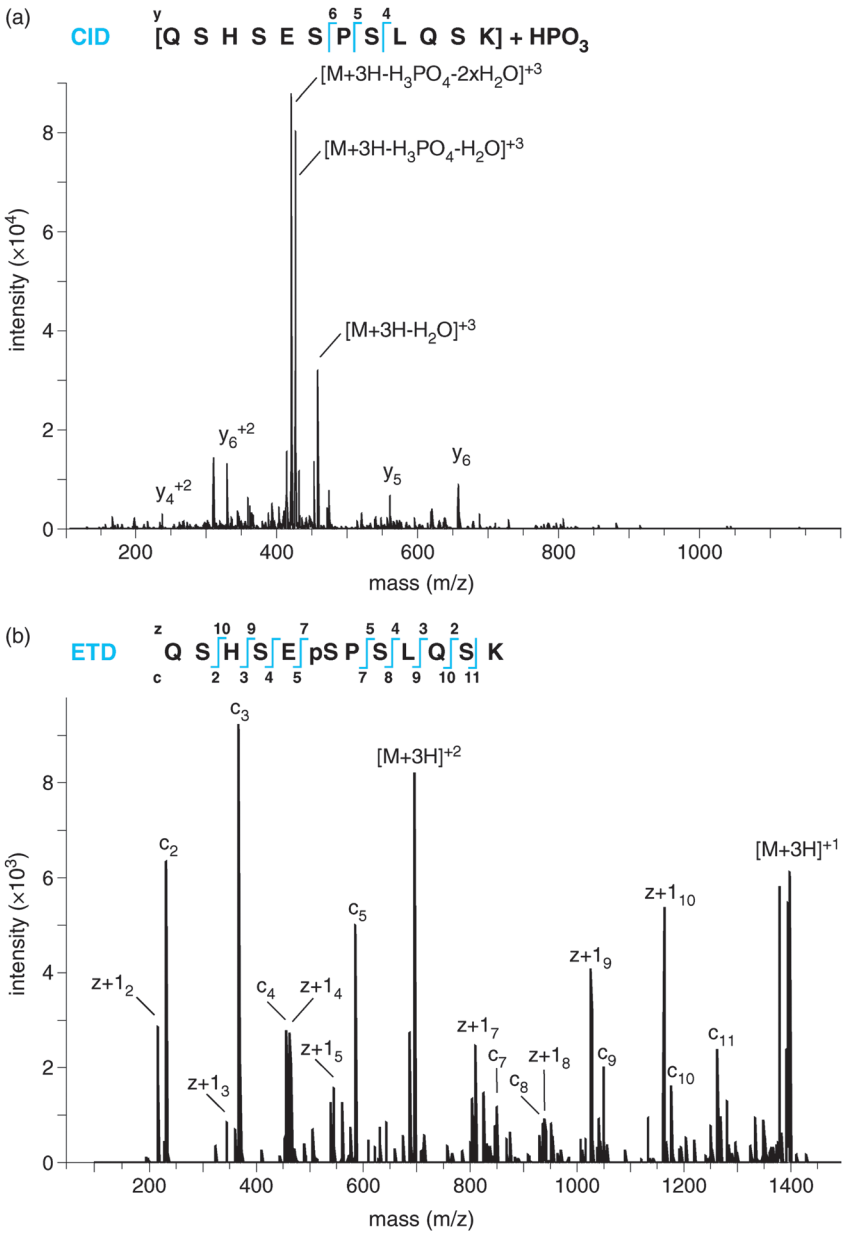


Figure 25.11 Comparison of CID (a) and ETD (electron transfer dissociation) (b) fragment mass spectra of the peptide QSHSESPSLQSK from human SRRM2 protein. The spectrum generated by CID exhibits dominant neutral loss fragments and little sequence specific fragments. In addition, the phosphorylated amino acid cannot be determined. In contrast, ETD fragmentation results in an almost complete c- and z-ion series and allows the exact localization of phosphoserine within the peptide. Source: courtesy of Heike Piechura, Freiburg, Germany.

sequentially fragmented in Q2 using CID. The resulting fragments are analyzed in Q3, which scans for fragment ions exhibiting a specific neutral loss like, for example, H_3PO_4. The detector records only those fragment ions that lose a specific neutral ion during collision in Q2. To obtain the sequence of the modified peptide it is subsequently fragmented and analyzed analogously to the product ions scan. In a neutral loss scan of phosphorylated peptides, phosphoserines and phosphothreonines exhibit a loss of H_3PO_4, which results in a mass shift of 98 Da (singly charged), 49 Da (doubly charged), and 32.7 Da (triply charged). Phosphotyrosines and acetyllysines cannot be characterized by means of neutral loss scan – because of their stability under the applied conditions they show no characteristic neutral loss.

In contrast, the PTM-specific immonium ion scan is applicable for peptides containing phosphotyrosines and acetyl-lysines. This method is based on a precursor ion scan, wherein Q3 scans for specific immonium and marker ions obtained from phosphotyrosines (216.042 Da) and acetyl-lysines (143.118 Da). Due to the loss of NH_3, a further specific marker ion for acetyllysine with a mass of 126.092 Da is generated. All three ions can be detected with high-resolution quadrupole-TOF hybrid mass spectrometers (phosphotyrosine by specific

immonium ion scan, acetyl-lysine specific by immonium ion scan). The main limitation of a PTM-specific immonium ion scan is the resolution and mass accuracy of applied mass spectrometers, because in the mass range of immonium ions various internal fragments with small mass differences are generated. To obtain the sequence of the modified peptide, and to determine the position of the phosphotyrosine or acetyl-lysine, the specific fragment ion spectrum is subsequently analyzed.

Fragmentation by CID is the most common method for *de novo* sequencing of peptides. However, it has limitations concerning the analysis of labile post-translational modifications, such as phosphorylation or glycosylation. For example, fragmentation of serine or threonine phosphorylated peptides using CID leads to a break of the labile phosphoryl bond resulting in the loss of H_3PO_4 or HPO_3 (neutral loss; Figure 25.11). The generated MS/MS-spectra exhibit a dominant precursor ion with a neutral loss and therefore provide little information about the sequence of the peptide.

To obtain detailed sequence information of modified peptides using ion trap mass spectrometers, additional fragmentation methods can be applied. The multistage activation (MSA), also known as pseudo MS^3, produces combined spectra of MS^2 and MS^3 fragmentation. First, a peptide ion is isolated in the ion trap and fragmented by CID (MS^2). This is followed by a second fragmentation step of the generated neutral loss fragment ion (pseudo-MS^3) to obtain sequence specific backbone fragments of the peptide. In contrast to a conventional MS^3 experiment, no additional isolation of the fragment ion is applied between both fragmentation steps. This means that the product ions of the neutral loss fragment and the initial MS^2 product ions are combined in the ion trap, resulting in a combined spectrum.

As an alternative fragmentation technique to CID, electron transfer dissociation (ETD) using ion trap MS is applied for the analysis of PTMs. Using ETD, the side chain of amino acids and modifications remain intact. The fragmentation mechanism is based on transferring the electrons of a radical anion with a low electron affinity (transfer reagent) to the positively charged peptide ion. A commonly used transfer reagent is fluoranthene. The positively charged peptide ion and the radical anion of fluoranthene are consecutively introduced into the ion trap, where they are mixed and the reaction takes place. Thereby, an unpaired electron is transferred to the peptide, whereby it becomes unstable resulting in a break of the peptide's backbone. The transfer of an electron leads to a reduction of the charged state of the peptide ion and also of the fragment ions. The resulting fragment ion pattern is independent of the peptide length, amino acid composition, and the presence of a PTM. Fragmentation by ETD produces singly charged c- and z-ions, which allow determination of the sequence of the peptide and the position of the modification directly from the MS^2 spectrum (Figure 25.11b).

In MALDI-TOF mass spectrometers, the loss of H_3PO_4 and HPO_3 takes place in the field-free region (*post source*) as a result of metastable and/or collision-induced decays. In linear mode, the resulting fragments cannot be distinguished from the mass of the precursor ion. In the reflector mode, the fragments are detected as broad and non-isotopically resolved signals. This means that the presence of signals in the reflector spectrum, which are a result of metastable decays ($[MH–H_3PO_4]^+$ and $[MH\text{-}HPO_3]^+$), indicates a phosphorylated peptide. Both fragment ions $[MH–H_3PO_4]^+$ and $[MH–HPO_3]^+$ are not detected at −98 and −80 Da, respectively, in a reflector spectrum as expected. This is because the reflector of the mass spectrometer is calibrated accurately for ion acceleration with full energy. Both fragment ions have lower kinetic energy than the corresponding precursor ions and therefore these are detected at slightly lower *m/z* values.

25.6 Quantitative Analysis of Post-translational Modifications

The biological function of a protein depends not only on the type and position of its PTMs but also on the stoichiometry of the modifications. Therefore, quantitative analysis of the phosphorylation or acetylation of a particular amino acid in a protein by MS is an important part of the modification-specific proteomic analysis. The quantification of a PTM can be carried out relatively between multiple conditions of a cell or a tissue, or absolutely based on internal standards. For relative quantification metabolic labeling or chemical labeling (using stable

isotopes) can be applied. Here, the proteins obtained under different conditions are labeled by stable isotopes, which differ slightly in their molecular mass, or by well-defined reporter groups.

Metabolic labeling is carried out by the addition of labeled amino acids in the medium during cell culture (stable isotope labeling by amino acids in cell culture, SILAC). Chemical labeling is carried out, for instance, using a combination of β-elimination and Michael addition, whereby a labeled thiol compound is added (Section 25.3). Thus, chemical labeling is only suitable for the analysis of serine and threonine phosphorylated proteins that are prone to β-elimination.

Proteome Analysis, Chapter 39

Commonly, thiol compounds containing an affinity tag (e.g., biotin) and an isotopically labeled linker (ICAT) are used for derivatization. This allows both the specific enrichment of phosphorylated proteins as well as MS-based relative quantification.

Absolute quantification (AQUA) of modified peptides can be carried out by the addition of internal isotopically labeled standard peptides. Initially, the sequence of the phosphorylated or acetylated peptide and the localization of the modified amino acid are identified by MS. Then, the identified peptide, which contains the modification, is chemically synthesized in the presence of stable isotopes, allowing the production of a peptide analog, which differs only in its mass from the naturally occurring counterpart. A defined amount of the synthetic peptide is then added to the protein sample as an internal standard. Subsequently, the proteins are proteolytically cleaved and analyzed by MS. Absolute quantification of the endogenous modified peptide is conducted by comparing the signal intensities of both peptide ions (endogenous and synthesized) in generated MS spectra (Figure 25.12).

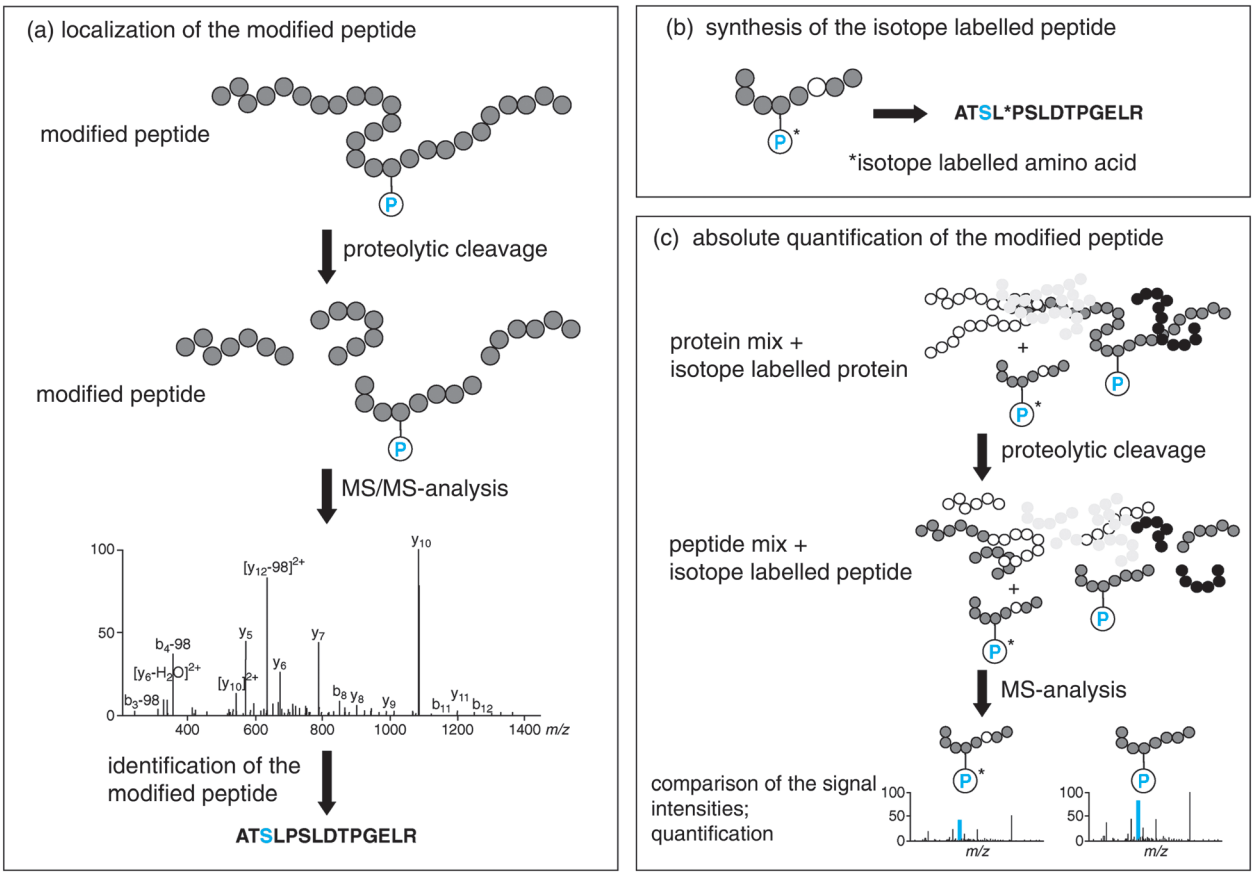


Figure 25.12 Absolute quantification of modified proteins and peptides using AQUA. The strategy for absolute quantification of modified proteins and peptides includes three steps: (a) the localization of the modified amino acid is determined by MS after proteolytic digestion or *in vitro* modification. (b) The corresponding modified peptide is synthesized in the presence of stable isotopes. This results in a peptide that is chemically identical to the naturally occurring modified counterpart. (c) The synthetic peptide is mixed with the sample in a defined amount as an internal standard. Both the native and the synthetic peptide are analyzed by MS. The intensities of selected peptide ions are calculated and compared with each other, whereby the absolute quantification of peptides is possible.

25.7 Future of Post-translational Modification Analysis

To increase our knowledge of the mechanistic details and of the biological functions of post-translational phosphorylation and acetylation in the regulation of protein functions and interactions, the qualitative and quantitative analysis of modified proteins has become more important for molecular biology as well as for proteome analytics. A proteome was originally defined as the protein complement of a genome, and due to its dependency on external and internal factors it is highly complex and dynamic. In addition to identification and quantification of individual proteins of a proteome, characterization of PTMs is the main task of proteome analysis. Besides the detection of numerous protein isoforms in high-resolution 2D-PAGE, PTMs connect the identified proteins of a cell or a tissue with functional data. Only with detailed knowledge of the modifications of proteins can the functional role of a biological system be deciphered. For the specific enrichment and analysis of PTMs various methods are available. However, the techniques must be carefully chosen and adjusted to the respective sample. Due to their specific limitations, many of the indicated methods can only be used in combination with each other for the successful characterization of phosphorylations or acetylations. Furthermore, a multidimensional and parallel purification of protein samples by LC, PAGE, and CE as well as optimization of the mass spectrometric methods is required. Therefore, it is expected that the number of qualitative and quantitative analyses of PTMs and the extent of modification-specific optimization will grow continuously in the future.

Further Reading

Boersema, P., Mohammed, S., and Heck, A. (2009) Phosphopeptide fragmentation and analysis by mass spectrometry. *J. Mass Spectrom,*. **44**, 861–878.

Dormeyer, W., Dorr, A., Ott, M., and Schnolzer, M. (2003) Acetylation of the HIV 1 Tat protein: an in vitro study. *Anal. Bioanal. Chem.*, **376**, 994–1005.

Choudhary, C., Kumar, C., Gnad, F., Nielsen, M.L., Rehman, M., Walther, T.C., Olsen, J.V., and Mann, M. (2009) Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science*, **325**, 834–840.

Farley, A. and Link, A. (2009) Identification and quantification of protein post-translational modifications. *Methods Enzymol.*, **463**, 725–763.

Jensen, O.N. (2004) Modification specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr. Opin. Chem. Biol.*, **8**, 33–41.

Salih, E. (2005) Phosphoproteomics by mass spectrometry and classical protein chemistry approaches. *Mass Spectrom. Rev.*, **24**, 828–846.

Sickmann, A. and Meyer, H.E. (2001) Phosphoamino acid analysis. *Proteomics*, **1**, 200–206.

Spange, S., Wagner, T., and Krämer, O.H. (2009) Acetylation of non-histone proteins modulates cellular signalling at multiple levels. *Int. J. Biochem. Cell Biol.*, **41**, 185–198.

Witze, E., Old, W., Resing, K., and Ahn, N. (2007) Mapping protein post-translational modifications with mass spectrometry. *Nat Methods*, **4**, 798–806.

Part IV

Nucleic Acid Analytics

Isolation and Purification of Nucleic Acids

26

Marion Jurk

Miltenyi Biotec GmbH, Friedrich-Ebert-Straße 68, 51429 Bergisch Gladbach, Germany

A basic and important prerequisite of nearly all nucleic analysis methods is pure and immaculate nucleic acid preparations. Without a faultless quality of the starting material all following methods will most likely fail. Possible contaminations in nucleic acid preparations are nucleases, nucleic acids of different kinds, macromolecules, or salts. Careful preparation of the nucleic acid material is therefore of special interest to any researcher. As the diversity of nucleic acids in nature is huge, the variety of methods of isolating them is similarly broad.

High molecular weight genomic DNA has to be isolated for obvious reasons by a different method than small, single-stranded RNA or circular plasmids. The organism or cells from which the nucleic acids are isolated also needs to be taken in account. Yeast cell walls have different lysis requirements than bacterial walls or cellular membranes. Dependent on the subsequent analysis methods, it could be sufficient to use a more timesaving but less pure purification method or it might be worth the effort to isolate very pure and high quality nucleic acids. For several years, the automated purification of nucleic acids has become increasingly important, for example, for diagnostic purposes or for genome sequencing. Huge numbers of nucleic acids are isolated in parallel using automated liquid handling systems. All these different requirements can only be accommodated by a huge variety of isolation and purification protocols. In this chapter, different isolation and purification methods will be presented, structured according to the kind of nucleic acid to be prepared. As almost all isolation methods have the final purification and concentration principles in common, these general principles of nucleic acid purification will be discussed in the first part of this chapter.

26.1 Purification and Determination of Nucleic Acid Concentration

26.1.1 Phenolic Purification of Nucleic Acids

Protein containing contaminations of nucleic acid preparations can be removed by extraction of the nucleic acid solution with buffered phenol. Proteins are denatured by phenol and can be found as an insoluble fraction in the so-called interphase between the aqueous phase and the organic phase. Some proteins, however, might also dissolve in the organic phase. For the purification of DNA the phenol needs to be buffered with Tris-HCl or TE (Tris-HCl/EDTA), pH 7.5 or 8.0. DNA is in part soluble in phenol if the phenol is not buffered with TE. Oxidation products of phenol can damage the nucleic acids; only distilled phenol of high quality should be used. Since phenol is a very toxic compound, special precautions should be taken regarding its handling. Many companies offer buffered phenol of molecular biology grade.

Beside extraction with pure phenol solutions, it is also common to carry out the extraction with phenol/chloroform/isoamyl alcohol (PCIA). The composition of buffered phenol and

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

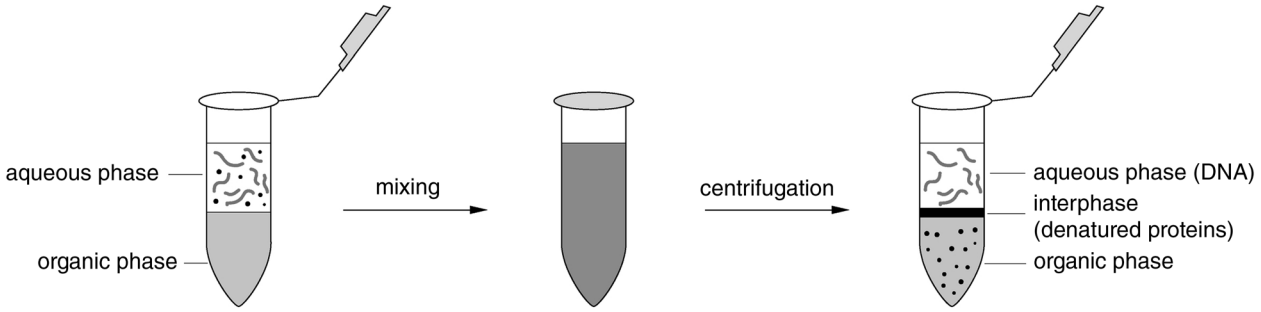


Figure 26.1 Phenolic extraction of aqueous DNA solutions. Protein containing decontaminations are removed by extraction with buffered phenol. Under common salt conditions, the aqueous phase is above the organic phase. The organic phase can be pure phenol equilibrated with Tris-HCl/EDTA or a mixture of phenol, chloroform and isoamyl alcohol. The denatured proteins are concentrated after phase separation in the so-called interphase. A minor part is also solubilized in the phenolic phase. The last step of the extraction is the extraction with chloroform/isoamyl alcohol.

chloroform/isoamyl alcohol is 25 : 24 : 1. Chloroform is a denaturing reagent and stabilizes the phase interface between the aqueous DNA solution and the phenol phase. There are proteins, for example, RNases, that cannot be denatured completely by phenol but can be by a mixture of phenol with chloroform. RNA molecules with long adenine-rich stretches (e.g., mRNAs) are especially soluble in pure phenol. By use of a phenol/chloroform mixture, the percentage of water solved in phenol will be reduced, accounting for a higher yield of nucleic acids. Isoamyl alcohol is added to reduce foam formation.

During phenolic extraction, the aqueous and organic phases are mixed thoroughly. Phase separation is then accelerated by centrifugation (Figure 26.1). The organic phase is in most cases at the bottom of the tube, with the aqueous phase on top. Denatured proteins are concentrated in the interphase. Phase inversion occurs if the nucleic acid solution contains high concentrations of salt (>0.5 M) or saccharose (>10%).

Phenolic extraction should be repeated to obtain complete denaturation of the proteins until no interphase is detectable. Since phenol is in part soluble in water, the phenol dissolved in the aqueous phase is removed by extracting the aqueous phase with a chloroform/isoamyl alcohol mixture.

For purification of RNA, the phenol should be buffered in water or low pH buffers ("acidic phenol"). DNA contaminants are much more soluble in acid phenol and can be removed more efficiently.

The nucleic acids purified by phenolic extraction can be precipitated by ethanol as described below (Section 26.1.3).

26.1.2 Gel Filtration

Gel Filtration/Permeation Chromatography, Section 10.4.1

Gel filtration methods can also be used for the purification of DNA (and RNA) solutions (the most common are Sephadex G50 or G75 and Bio-Gel P2). The purification effect is based on size exclusion allowing the separation of certain nucleic acid contaminations. Huge DNA molecules elute much faster (usually in the void volume of the column) than smaller, low molecular weight contaminants that become trapped in the pores of the column material and therefore elute at later time-points of the chromatography.

The nucleic acid containing solution is loaded onto the gel filtration column and the column is eluted with buffer. The eluate is collected in fractions and can be tested for nucleic acid content. Gel filtration columns can be self-made (Figure 26.2) using a glass Pasteur pipette, which can be sealed with glass wool or a glass beads. When purifying only very small amounts of DNA, the glass pipette should be treated with silane prior to purification as DNA sticks to the glass. Many companies offer suitable gel filtration or purification columns at relatively low cost. Loading volume and elution volume are predefined by the supplier. In contrast to regular gel filtration columns, spin columns use centrifugal forces to apply and elute the fractions, and are therefore much faster.

Reversed Phase Chromatography, Section 10.4.2

Instead of gel filtration, the principle of reversed-phase chromatography can also be applied. The nucleic acid solution is loaded onto the column material at low salt concentrations and is eluted with high salt buffer (e.g., Elutip columns).

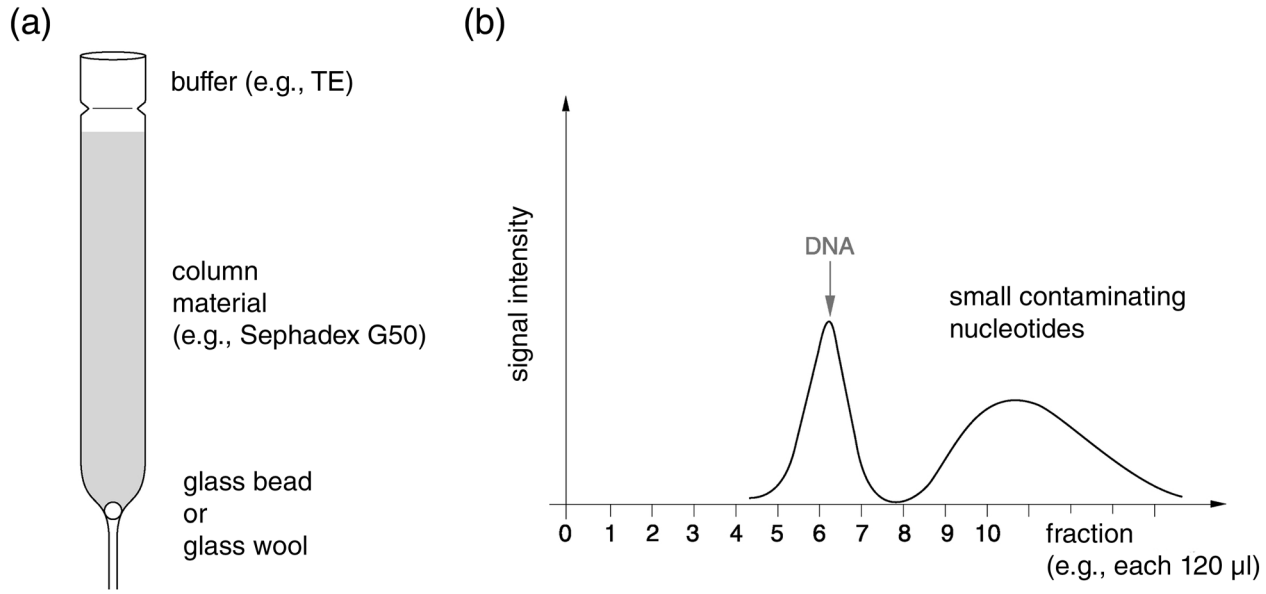


Figure 26.2 Gel filtration for purification of DNA solutions. (a) Gel filtration columns can be made out of Pasteur pipettes and are filled with equilibrated column material. Depending on the size of the DNA, Sephadex G50, Sephadex G25, or Sephacel materials are used. (b) Using the molecular sieve effect small molecules, contaminating nucleotides, or salts are withheld, while large DNA molecules are not detained and will elute first from the column. A tentative column profile is depicted. The fractions containing the DNA can be analyzed by OD determination, ethidium bromide staining, or when purifying radioactive DNA by radiation analysis. The maxima are closer together when small DNA molecules are purified.

26.1.3 Precipitation of Nucleic Acids with Ethanol

The most common method for the concentration and further purification of nucleic acids is by precipitation with ethanol. In the presence of monovalent cations, DNA or RNA form an ethanol-insoluble precipitate that can be isolated by centrifugation.

Monovalent cations are usually supplied by addition of sodium acetate or ammonium acetate. Ammonium acetate is used to reduce co-precipitation of free nucleosides. However, ammonium ions can inhibit the activity of certain enzymes, for example, T4 polynucleotide kinase.

For some applications, RNA is precipitated in the presence of lithium chloride. Lithium chloride is soluble in ethanol and will not be precipitated together with the nucleic acids. Chloride ions can act as inhibitors for several reactions, therefore precipitation with chloride ions should only be used in certain circumstances.

In the laboratory, the nucleic acid solution is adjusted to the desired salt concentration using a higher concentrated stock solution of, for example, sodium acetate. To this solution the 2.5–3-fold volume of ethanol is added and incubated, depending on the nucleic acid to be precipitated, at room temperature or −80 °C and centrifuged (Figure 26.3). The precipitated salt can be

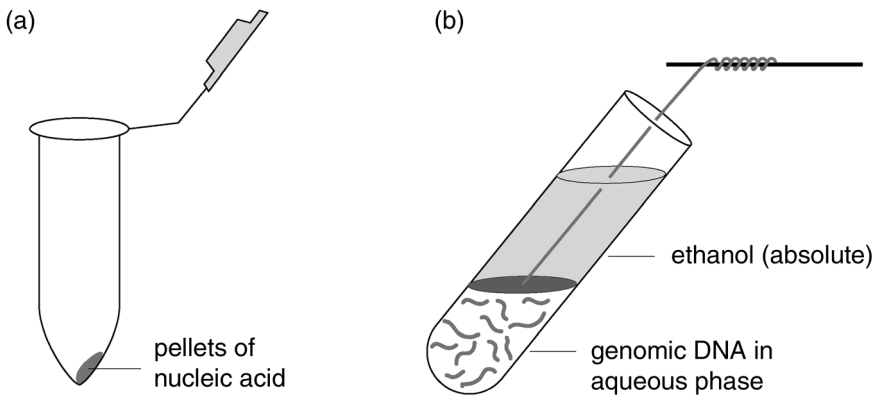


Figure 26.3 Precipitation of nucleic acids with ethanol. (a) To the aqueous nucleic acid solution a 2.5-to 3-fold volume of absolute ethanol (or in the case of isopropanol 0.5–1-fold volume) is added and the nucleic acids are precipitated by centrifugation. A colorless pellet is usually visible on the bottom of the tube. (b) High molecular weight genomic DNA can be precipitated on the phase interface by cautious overlay of the aqueous phase with ethanol. The DNA can be visualized by winding it on a sterile rod.

removed by washing the pellet with 70% ethanol. In contrast to the precipitated DNA, most salts are soluble in 70% ethanol. The nucleic acid pellet is dried briefly and re-solved in buffer or water.

Nucleic acids can also be precipitated by the addition of 0.5–1 volume of isopropanol. This protocol is advantageous if the volume of the reaction should be kept to a minimum. Sodium chloride is precipitated better in isopropanol. Isopropanol is less volatile than ethanol; the nucleic acid pellet needs to be washed carefully with 70% ethanol.

The carrier has to be chosen so that the material does not interfere with the subsequent reactions or applications. For example, tRNA will also be phosphorylated by T4 polynucleotide kinase and should not be used as carrier if the subsequent reaction involves phosphorylation reactions. Glycogen can interact with DNA–protein complexes.

Precipitation of Small Amounts using Carrier Material As very low concentrations of RNA or DNA (<10 µg ml^{-1}) are not efficiently precipitated, carrier material can be added prior to precipitation. The carrier material could be tRNA, glycogen, or linear polyacrylamide, all of which are commercially available. Carriers are also precipitated by ethanol and support the precipitation of low concentrations of nucleic acid.

26.1.4 Determination of the Nucleic Acid Concentration

Absorption Measurement, Section 7.1.3

The determination of the concentration of a nucleic acid solution is based on the characteristic absorption maximum at 260 nm (Figure 26.4). The absorption is caused by the aromatic rings of the bases. The absorption is measured photometrically at 260 nm using a quartz cuvette. Quartz does not absorb light of this wavelength. Most quartz cuvettes possess a path length of 1 cm. A solution containing 50 µg ml^{-1} double-stranded DNA has an absorption value of 1, called the optical density (OD).

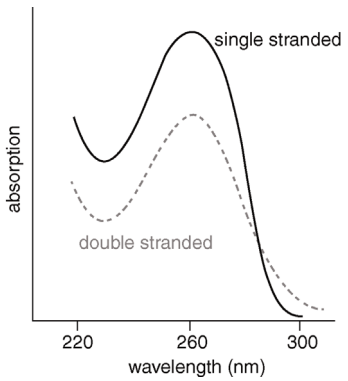


Figure 26.4 Absorption curves of double- and single-stranded DNA. The absorption maxima of nucleic acids is at 260 nm. The hyperchromic effect is shown, that is, the increase in the extinction during the transition of double- to single-stranded DNA.

The OD value is used to determine the concentration of a nucleic acid solution. Single-stranded nucleic acids have a higher absorption, an effect called hyperchromicity. Consequently, different values need to be used when calculating single- or double-stranded RNA or DNA solutions (Table 26.1). For simultaneous concentration measurements of a huge number of nucleic acid solutions in parallel, a quartz 96-well plate can be used with specially equipped reader instruments. In these instruments, the path length needed for the concentration calculation is determined automatically.

To determine the concentration of short, single-stranded oligonucleotides of known sequence, other values are used for calculation. The absorption coefficient of each base is summed up for the oligonucleotide accordingly. This absorption value equals an oligonucleotide concentration of 1 µmol ml^{-1}. For very exact determination of the absorption value of oligonucleotides, the nearest neighbor method is used, which accounts also for the sequence of the oligonucleotide.

Table 26.1 Photometric determination of the concentration of nucleic acid solutions. The concentration of oligonucleotides can be estimated using the value for single-stranded DNA or, if the sequence of the oligonucleotides is known, by using the sum of molar extinction coefficients of the individual bases. Source: According to: Maniatis, T., Fritsch, E.F., and Sambrook, J. (1989) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor.

	Concentration corresponding to 1 OD260 (µg ml^{-1})
Double-stranded DNA	50
Single-stranded DNA	40
Single-stranded RNA	33
	Molar extinction coefficient of single nucleotides ε (mM^{-1} cm^{-1})
(dATP)	15.4
(dCTP)	9.0
(dGTP)	13.7
(dTTP)	10.0

The absorption maximum of proteins is based on the absorption of the aromatic amino acids at 280 nm. By determination of the absorption at 260 and 280 nm, it is possible to estimate the purity of the nucleic acid solution. A pure DNA solution has a ratio $OD_{260\ nm}/OD_{280\ nm}$ of 1.8, for a pure RNA solution the ratio is 2.0. The value is considerably lower if the nucleic acid solution is contaminated with proteins. A solution containing 50% protein/DNA has a ratio of approx. 1.5.

Determination of the OD of a nucleic acid solution usually requires concentrations higher than 25 µg ml^{-1}. Nucleic acid solutions with a low concentration can be estimated using ethidium bromide staining and comparison with dilutions of known concentrations. A rough estimate of the concentration is also possible using the detection limit of agarose gels (approx. 20 ng).

A very sensitive quantification method is staining of the DNA with Hoechst 33258. Hoechst 33258 is a fluorescent dye of the bisbenzimide group and intercalates to the A/T rich regions of the DNA. The dye does not stain RNA, allowing for the quantification of DNA solutions contaminated with RNA. Since the A/T content influences the result of the quantification, a quantification standard should contain similar amounts of A/T. For determination of genomic DNA, calf thymus DNA is used as standard with an A/T content of approx. 58%.

26.2 Isolation of Genomic DNA

Genomic DNA can be obtained from various sources. The isolation of genomic DNA from tissue, cell culture, plants, yeast, or bacteria is performed using a similar method, adapted to species (Table 26.2). Genomic DNA is a high molecular weight DNA that can be easily broken into smaller pieces by shear forces. Genomic DNA solutions need to be handled with caution. Pipetting through cannulas or pipette tips with small diameters should be avoided. As ethanol precipitation also has a negative impact on molecule size, high molecular genomic DNA (>150 kb) is not precipitated but purified using dialysis or extraction with 2-butanol.

Extremely high molecular weight DNA can be analyzed by melting the cells or tissue to be analyzed into agarose and all the following steps are then performed in these agarose blocks.

Lysis of Cell Membranes and Protein Degradation A fundamental step during isolation of genomic DNA is the proteolysis of cellular proteins by proteinase K. Simple extraction of the proteins using phenol is not sufficient. In addition, genomic DNA is complexed with histones and histone-like proteins, which cannot be removed completely by phenolic extraction. The optimal incubation temperature for proteinase K is between 55 and 65 °C. The enzymatic performance is optimal at 0.5% sodium dodecyl sulfate (SDS). Incubation of the starting material with proteinase K containing buffer is in many cases sufficient to disrupt and lyse the cell membranes. The addition of RNAse removes the contaminant RNA.

In many cases, the cells or tissue need to be disrupted mechanically before protease digestion. Homogenizers disrupt the cells with blades that rotate at high frequencies. A so-called French press or ball mills can also be used. A French press uses high pressure to disrupt the cells. Ball mills contain very fast moving small beads made of glass or steel. If the DNA is obtained from tissue, the tissue is shock frosted using liquid nitrogen and then pulverized to achieve a homogenous mixture.

The lysis of bacterial walls is achieved by lysozyme and yeast cell walls are disrupted by zymolase or lyticase specifically degrading yeast cell walls. The enzymes are inactivated during protease K treatment.

Table 26.2 Enzymes and lysis reagents for isolation of genomic DNA.

Nucleic acid origin	Lysis by	Subsequent treatment
Eukaryotic cell cultures	Sodium dodecyl sulfate (SDS)	Proteinase K
Tissue	Sodium dodecyl sulfate/proteinase K	Proteinase K
Plants	SDS or *N*-laurysarkosin	Proteinase K
Yeast (*Saccharomyces cerevisiae*; *Schizosaccharomyces pombe*)	Zymolyase or lyticase	Proteinase K
Bacteria (*Escherichia coli*)	Lysozyme	Proteinase K

Purification and Precipitation of Genomic DNA The proteinase K is inactivated and removed by phenolic extraction. The precipitation of genomic DNA after addition of ethanol can be easily observed: the DNA precipitates at the interphase between the water and ethanol and can be rolled-up as filaments on a sterile rod (Figure 26.3). Genomic DNA should be dried very cautiously to avoid solubility problems later on. The genomic DNA can be dissolved by incubation for several hours at 4 °C. Genomic DNA isolated by ethanol precipitation is of sufficient purity for most applications.

Phenolic extraction and subsequent ethanol precipitation will result in an average molecular size of around 100–150 kb. This size is sufficient for the generation of DNA libraries using the bacteriophage λ-vectors and for Southern blot analysis. The construction of cosmid libraries requires DNA fragments of at least 200 kb, and so extraction with organic solvents cannot be used. The proteinase K and remaining proteins are denatured using formamide and removed by dialysis using collodion bags. This method avoids shear forces and results in high molecular DNA fragments (>200 kb).

A fast and easy isolation method is the lysis of cells and denaturation of the proteins with guanidinium hydrochloride. The DNA is isolated by ethanol precipitation. This method yields genomic fragments with an average size of 80 kb and can be used for Southern blot or PCR analysis.

Additional Purification Steps Genomic DNA can be purified using CsCl gradient centrifugation. During centrifugation RNA contaminations are pelleted and removed completely. The CsCl density gradient centrifugation is described in Chapter 1 and in Section 26.3.

Commercially available kits use the anion exchange method for the isolation and purification of genomic DNA. The kits are available with columns containing the anion exchange material using gravity or spin-columns using centrifugal forces. This purification method does not need any organic extraction; however, the DNA is subject to shear forces so that high molecular weight DNA cannot be isolated using this method. The genomic DNA isolated by the anion exchange columns can be used for Southern blot, PCR, and next generation sequencing.

Polysaccharide contaminants can be removed by treatment with CTAB (cetyltrimethylammonium bromide, Figure 26.5). This purification step is essential when genomic DNA is isolated from plants or bacteria, as they contain high levels of polysaccharides. CTAB complexes polysaccharides and removes the remaining proteins. By addition of chloroform/isoamyl alcohol the complexed polysaccharides are precipitated in the interphase. An important factor is the NaCl concentration: if the concentrations is below 0.5 M, the genomic DNA will also precipitate in the presence of CTAB.

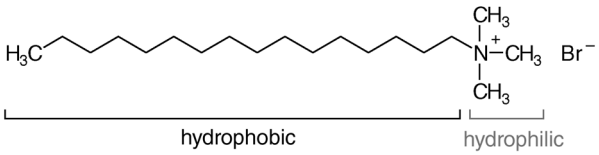


Figure 26.5 CTAB (cetyltrimethylammonium bromide or hexadecyltrimethylammonium bromide). The quaternary ammonium salt can act as cationic detergent.

26.3 Isolation of Low Molecular Weight DNA

26.3.1 Isolation of Plasmid DNA from Bacteria

Plasmids, that is, extrachromosomal mostly circular DNA occur naturally in microorganisms. Plasmids consist of 2 to more than 200 kb and fulfill various different genetic functions. In daily laboratory business, plasmids consist of defined genetic elements (replication origin, resistance gene, and polylinker – Figure 26.6). These so-called plasmid vectors are essential tools for a huge variety of applications.

The methods described below deal exclusively with the isolation of bacterial plasmid vectors. Plasmids are grown in bacteria using antibiotic selection. The plasmids contain at least one selection gene for resistance to certain antibiotics, for example, the bla-gene coding for β-lactamase enables bacteria that carry the gene to grow in ampicillin-containing media. In addition, plasmids contain a bacterial origin of replication for propagation of the plasmid in the bacterium. The kind of origin of replication determines the copy number of a plasmid in the bacterium (Table 26.3.).

Table 26.3 Origins of replication of commonly used plasmid vectors and copy numbers.

Plasmid	Origin of replication	Gene of resistance	Copy number
pBR322 and derivatives	pMB1	Amp^r, Tet^r	15–20
pUC	pMB1	Amp^r	500–700
pBluescript	pMB1	Amp^r	300–500
pGEM	pMB1	Amp^r	300–400
pVL 1393/1392	ColE1	Amp^r	>15
pACYC	p15A	$Chloramphenicol^r$, Tet^r	10–12
pLG338	pSC101	Kan^r, Tet^r	ca. 5

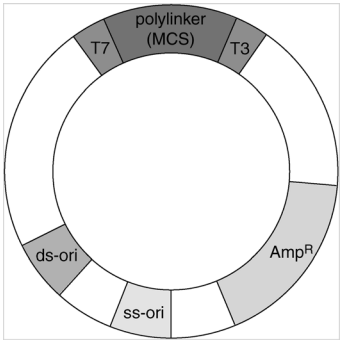


Figure 26.6 Composition of a typical plasmid vector for cloning and amplification of DNA fragments. The fragment of interest is cloned to the artificial multiple cloning site (MCS). T7 and T3 are promoters that are recognized specifically by RNA polymerases of the T7 and T3 bacteriophages and are used for RNA synthesis of the cloned fragments. Amp^R depicts a gene for selection that renders bacteria containing the vector resistant to ampicillin: the bacteria can grow in ampicillin containing medium. The replication of origin (ori) is necessary for the autonomous replication of the plasmid. The ori region enables double-stranded replication of the plasmid whereas a second origin of replication (e.g., f1 single-stranded phages) permits single-strand replication (ss-ori).

Plasmids are classified as low copy (copy number < 20) and high copy plasmids (copy number > 20). The copy number of the plasmids is a major determinant for the yield of a plasmid from a bacterial culture. Most plasmids contain a mutated version of the pMB1 origin of replication derived from ColE1 multi-copy plasmids of the Enterobacteriaceae family.

The isolation steps can be divided into growth and lysis of the bacteria and isolation and purification of the plasmid DNA.

Bacterial Culture For the isolation of plasmids, derivatives of the *Escherichia coli* strain K12 are used. The strain is considered biologically safe as it is missing pathogenic genes (e.g., factors relevant for adhesion and invasion, toxins, and certain surface molecules). Not all *Escherichia coli* K12 strains are equally useful for plasmid production. Good host strains are, for example, DH1, DH5a, and XL1 Blue. Certain strains like HB101 and JM100 express a high amount of endonucleases and carbohydrates that are detrimental to the plasmid isolation procedure. To avoid mutations and unwanted DNA recombinations, strains deficient of recombinase A (recA-), like XL1 Blue and its derivatives, are preferred.

Bacteria are grown in liquid culture using autoclaved Luria-Broth (LB, contains yeast extract, Bacto tryptone, and sodium chloride) in the presence of the antibiotic. The amount and quality of the added antibiotic is important for the plasmid yield. Ampicillin is temperature sensitive and should not be added to hot autoclaved medium.

According to good microbiological practice, the broth is inoculated with a single bacterial colony, which is first grown in a small volume of media and then diluted to the needed volume. According to the amount of culture volume the DNA preparations are classified as "mini-" (1–10 ml), "midi-" (25–100 ml), or "maxi-" (>100 ml) preparations.

The yield of low copy plasmids can be increased by the addition of chloramphenicol to the media (see below). Chloramphenicol is also used for the isolation of high copy plasmids as it keeps the number of bacteria and thus the amount of bacterial debris low.

Plasmid containing a ColE1 origin of replication can be amplified selectively compared to the bacterial genome. During the logarithmic growth phase an inhibitor of translation (e.g., chloramphenicol) is added to the bacterial culture. Chloramphenicol inhibits the synthesis of the Rop (repressor of primer) protein. This protein accounts for the control of the copy number of plasmids. Inhibition of this protein results in an increased replication of the plasmid (relaxed replication).

Lysis of Bacteria Many different methods are available for the lysis of the plasmid containing bacteria (Table 26.4). The method of choice depends on the type and use of the plasmid to be

Table 26.4 Methods used to reveal bacteria.

Method of digestion	Means of analysis	Comment
Alkaline	SDS/NaOH	Quick and easy. The most suitable method for large plasmids and low-copy plasmids
Koch lysis	Lysozyme/100 °C	Endonuclease A is inactivated completely
Lithium method	LiCl/Triton X-100	Quick and efficient, not suitable for large plasmids (>10 kb)
SDS lysis	Lysozyme/SDS	Frequently used for large plasmids (>15 kb)

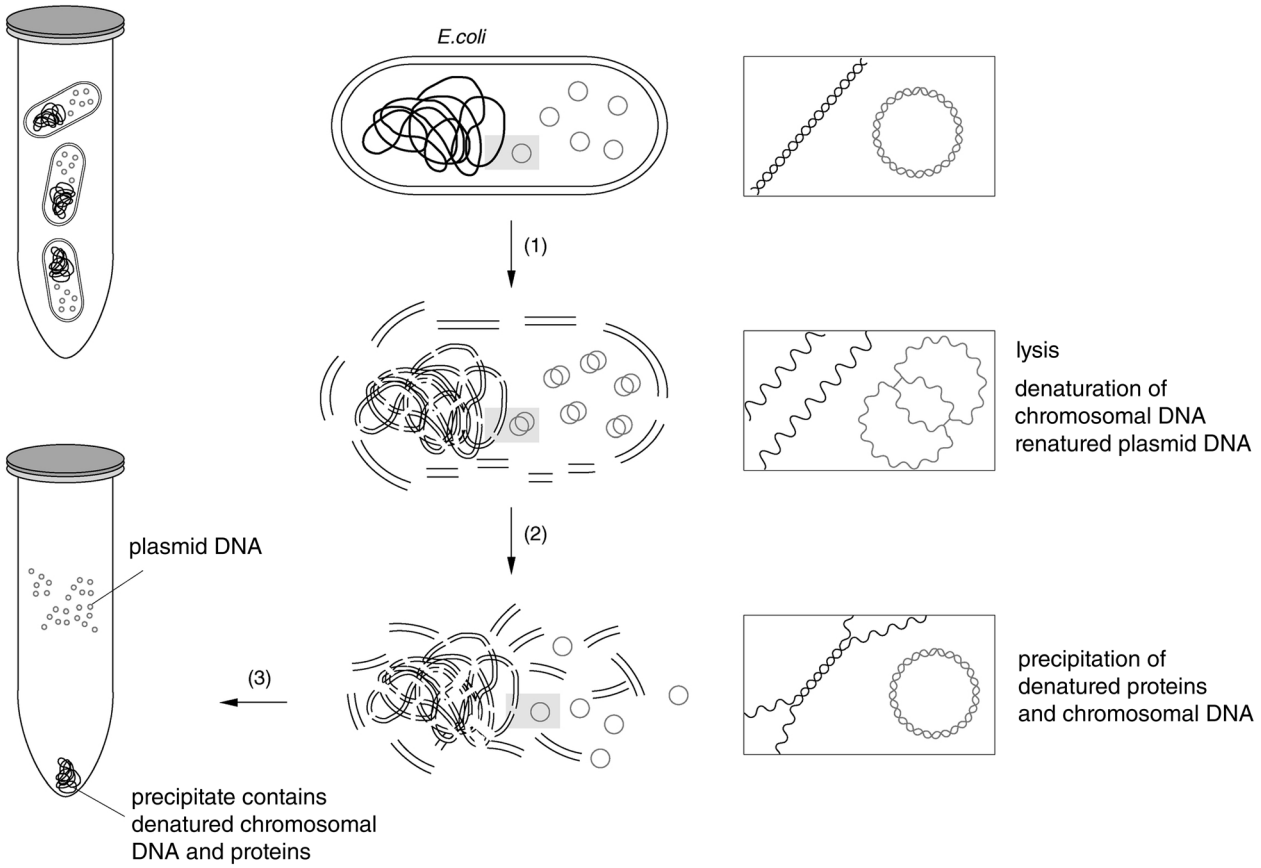


Figure 26.7 Principle of the alkaline lysis of bacteria for the isolation of plasmid DNA. (1) The bacteria are lysed using SDS and the DNA is denatured by NaOH. (2) The solution is neutralized by the addition of sodium acetate. Denatured proteins and chromosomal DNA are precipitated together with the potassium salt of the dodecyl sulfate. Low molecular plasmid DNA remains in solution and renatures. (3) The insoluble complexes are separated by centrifugation and the plasmid DNA can be isolated. Source: adapted according to: Micklos, D.A. and Freyer, G.A. (1990) *DNA Science. A First Course in Recombinant DNA Technology*, Cold Spring Harbor Laboratory Press and Caroline Biological Supply Company, Cold Spring Harbor.

isolated. The most common method is alkaline lysis (Figure 26.7). The bacterial culture is centrifuged and the pellet resuspended in a buffer containing EDTA. EDTA complexes bivalent cations (Mg^{2+}, Ca^{2+}) that are important for the structural integrity of the bacterial walls. The buffer can also contain RNase A to degrade most of the bacterial RNA at this first step.

It is important that the RNase A does not contain any contaminating DNases. This can be achieved by incubation of the RNase A solution at 95 °C. RNase H is a very stable enzyme that renatures after heat treatment to yield an active enzyme whereas DNases are permanently inactivated.

The bacterial suspension is lysed completely by the addition of SDS and NaOH. SDS functions as detergent, solubilizing the phospholipids and proteins of the bacterial cell walls. Sodium hydroxide denatures proteins, chromosomal, and plasmid DNA. The timespan of the incubation of the solution under alkaline conditions is important for the quality of the plasmid DNA. Too long an incubation time leads to irreversible denaturation of the plasmid; too short an incubation results in incomplete lysis of the bacteria and low plasmid yield. Completely denatured plasmid DNA can be detected using gel electrophoresis on an agarose gel. Denatured plasmid DNA has a higher mobility than superhelical plasmid DNA and is stained less by ethidium bromide.

The lysate is neutralized with potassium acetate buffer. Potassium dodecyl sulfate has a much lower solubility in water than sodium dodecyl sulfate and precipitates at high salt concentrations present in the lysate. Denatured proteins, high molecular weight RNA and denatured chromosomal DNA, and cellular debris form insoluble complexes in the presence of potassium dodecyl sulfate and will be co-precipitated with the potassium dodecyl sulfate. The smaller plasmid molecules remain in solution and renature upon neutralization of the solution. The insoluble debris is centrifuged and the supernatant can be processed further. For some applications the purity of the DNA solution is sufficient and the plasmid DNA can be precipitated with ethanol or isopropanol and washed with 70% ethanol.

This quick and easy method is useful for the preparation of many plasmids simultaneously and is used to check cloning efficiency. Many different single bacteria colonies are inoculated in a small volume of media and successful cloning is checked by digestion of the plasmid DNA with restriction enzymes if the plasmid contains the desired insert.

Commercially available kits are based on the alkaline lysis principle. The plasmid DNA is purified as described below by anion exchange chromatography before precipitation with ethanol.

As well as alkaline lysis, bacteria can be lysed thermally (boiling lysis). The bacterial cell walls are broken down by addition of lysozyme and lysed bacteria are heated for a short time. The debris is pelleted and the plasmid DNA can be isolated by ethanol precipitation. This method does not inactivate completely endonuclease A present in some *E. coli* strains (HB101, *endA*+). The plasmid DNA should therefore be purified by phenolic extraction prior to precipitation.

Other (more uncommon) methods are incubation with the non-ionic detergent Triton X100 (Figure 26.8) in the presence of lithium chloride or lysis by SDS and lysozyme. The latter method (without the addition of NaOH) is used when high molecular plasmids need to be isolated. High molecular weight plasmids cannot be renatured completely in the presence of NaOH.

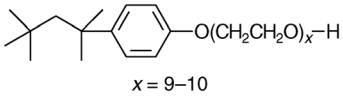


Figure 26.8 Non-ionic detergent Triton X-100.

Lysozyme: Abundant hydrolase found in saliva and tear fluid. Lysozyme hydrolyzes the 1,4-β-linkages between *N*-acetylmuramic acid and *N*-acetyl-D-glucosamine present in bacterial cell walls.

Purification of DNA by Anion Exchange Chromatography In general, commercially available columns are used for purification by anion exchange. The positive charge is provided by protonated diethylammoniumethyl (DEAE) groups. The negatively charged DNA is bound at lower salt concentrations (750 mM) to the column material. Proteins and degraded RNA are not bound to the column material under these conditions. The column material is washed with buffer containing a higher salt concentration (1 M) to elute traces of bound protein or RNA. DNA does not elute from the column under these conditions. The DNA is eluted at even higher salt concentrations (1.25 M). The exact buffer conditions are dependent on the column material and supplier. Table 26.5 depicts an overview of expected yields using anion exchange columns.

Anion Exchange Chromatography, Section 10.4.7

Several purification protocols have been set up for the removal of endotoxins prior to the purification by anion exchange. The lipopolysaccharides adherent to the bacterial membranes are treated with detergents (*n*-octyl-β-D-thioglucopyranoside, OSPG) to remove binding proteins. Then, the lipopolysaccharides are removed using columns loaded with polymyxin B. This antibiotic binds lipopolysaccharides very efficiently. Ultrapure DNA with very low toxin content can be obtained by repeated purification using a CsCl density gradient.

With the described purification method it is possible to co-purify certain lipopolysaccharides present in almost all Gram-negative bacteria. The presence of these so-called endotoxins is critical when transfecting DNA in sensitive cells or cell lines. Endotoxins can reduce transfection efficiency and can result in stimulation of protein synthesis or activation of the innate immune or the complement system.

Purification of DNA by Density Gradient Centrifugation Ultrapure, high yield DNA can be obtained by centrifugation using a CsCl density gradient. Due to the significantly increased quality of the commercially available anion exchange kits, the method of density gradient centrifugation has lost its relevance and will therefore only be summarized briefly.

The isopycnic centrifugation of DNA molecules within a CsCl density gradient is performed in the presence of ethidium bromide. The mechanism and thermodynamic aspects of ethidium bromide intercalation will be discussed in Chapter 27.2. Plasmid DNA and chromosomal DNA

Table 26.5 Approximate DNA yield after anion exchange purification. The yield of high copy plasmids is approx. 2–5 µg ml^{-1}, of low copy plasmids 0.1 µg ml^{-1}.

Vector	Plasmid type	Bacterial culture (ml)	Yield (µg)
pUC, pGEM	High copy	25	50–100
pUC, pGEM	High copy	100	300–500
pBR322	Low copy	100	50–100
pBR322	Low copy	500	100–500

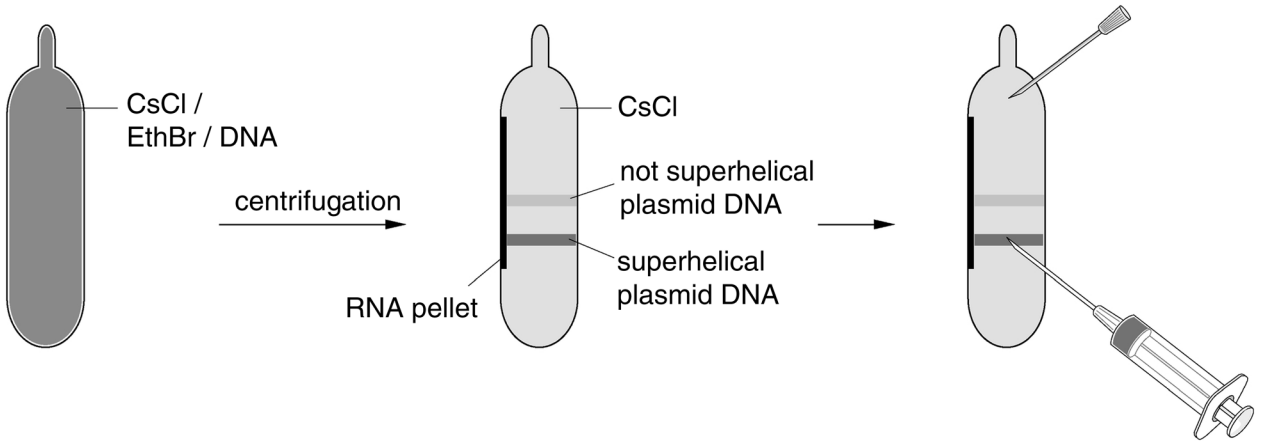


Figure 26.9 Purification of plasmid DNA by CsCl density gradient centrifugation in the presence of ethidium bromide. Using vertical rotors, the gradient is generated parallel to the axis of the rotor. After stopping the rotor and centrifugation, the gradient tips but keeps its layers so that the plasmid is visible as circular bands. Due to the different densities, superhelical form I and forms II and III (open form, linear form) can be separated within the gradient. RNA–ethidium complexes are pelleted on the wall of the centrifuge tube. To isolate the plasmid, the sealed tube is ventilated using a needle. The plasmid is obtained by puncturing the tube with a second cannula below the plasmid band. The DNA containing solution is aspirated with a syringe. If genomic DNA was present, it can be found above form I because of its lower density. For ultrapure DNA, the CsCl density gradient purification is repeated.

can be distinguished by their different densities with intercalated ethidium bromide. Ethidium bromide intercalates into double-stranded DNA, preferentially linear or nicked plasmid DNA, and to a lower extent into covalently closed-circular plasmid DNA. The resulting differences in density are used to separate the different molecular forms of DNA (Figure 26.9). The buoyant density of RNA is higher than the maximal density of CsCl, resulting in RNA/Ethidium bromide pellets. RNA separation is achieved using Cs_2SO_4.

The ethidium bromide is removed by repeated extraction of the DNA solution with *n*-butanol. Any remaining traces of ethidium bromide are completely removed by phenolic extraction. The high concentration of CsCl can be removed by dialysis of the DNA against TE buffer or water. The DNA can also be diluted to low concentrations and then precipitated using ethanol.

26.3.2 Isolation of Eukaryotic Low Molecular Weight DNA

Yeast Plasmids The isolation of ultrapure yeast plasmids is difficult. In practice, total DNA is isolated. Since yeast plasmid contains a yeast origin of replication and a bacterial origin of replication pure yeast plasmids are isolated after re-transformation into *E. coli* as contaminating chromosomal yeast DNA cannot be replicated in bacteria.

Hirt Extraction Low extrachromosomal DNA like plasmid DNA or viral DNA from cell or tissue cultures is isolated using a protocol established by B. Hirt 1967 for the isolation of polyomavirus DNA from murine cells. The cells are lysed using 0.5% SDS and adjusted to 1 M NaCl. The mixture is incubated over night at 0 °C and centrifuged. The supernatant contains low molecular weight DNA and can be purified using proteinase K and phenolic extraction.

26.4 Isolation of Viral DNA

26.4.1 Isolation of Phage DNA

Bacteriophage λ and others are widely used as vectors for phage display, as reporter phages, or for cloning of genomic libraries. No other cloning system allows insertion of high molecular weight DNA fragments (10–20 kb); it can also be used very conveniently for high throughput screening. It may be necessary to isolate and analyze the DNA fragment inserted into the phage genome.

Proliferation of Phages Bacteriophages are proliferated using liquid culturing of *E. coli*. The choice of host strain is dependent on the bacteriophage strain. The bacteria are grown in the log phase using maltose. Maltose induces expression of the bacterial receptor (*lamB*) for bacteriophage λ. The bacteria are harvested and the culture is adjusted to a certain density using a buffer containing Mg^{2+} (λ-diluent, SM media). The bacterial cell number is determined photometrically: absorption of the culture is measured at 600 nm (blank is pure media). 1 OD equals approx. 8×108 bacteria. The Mg^{2+} ions stabilize phage particles and Mg containing media is used for phage proliferation (NZCYM media).

For optimal proliferation of bacteriophages, the initial ratio of phages to bacteria is important. If the number of phages outweighs the number of bacteria in the initial phase, the bacteria are lysed completely and no phage proliferation can occur, and the yield will be very low. If the bacterial culture is infected initially with a low number of phages, the bacteria will overgrow a possible phage infection and the phage yield is also low. The initial ratio needs to be determined for phage and bacterial strain. A good ratio is found when complete lysis takes more than 8 h. Complete lysis of a bacterial culture can be detected by the sudden clearance of the turbid bacterial culture and the sudden occurrence of lysed bacterial debris.

DNA library (genomic or cDNA library): Genomic DNA libraries contain the whole genome of an organism but split into smaller fragments that can be handled and cloned. The genome is fragmented enzymatically and then cloned into suitable vector systems, like the bacteriophage λ genome. Beside the genomic libraries, cDNA libraries represent the mRNA spectrum of a cell or organism. The cDNA is generated by reverse transcription of the mRNA.

Isolation of Phage Particles To remove bacterial RNA and DNA, RNases and DNase is added to the bacteriophage cultures. The phage DNA will remain intact, protected by the intact phage capsid. The phage particles are isolated by ultracentrifugation (100 000*g*). The purity of the phages is, for most applications, sufficient. Phages form a colorless to light brown pellet that is resuspended in TE. Phage particles are very sensitive to complexing agents that decrease the Mg ion concentration. Resuspension of the pellet in EDTA containing buffers destabilizes the capsid and facilitates later lysis. In some protocols the particles are precipitated using poly(ethylene glycol) (PEG). If a higher purity of the phages is desired they can be purified by CsCl density gradient centrifugation. Isolated and purified phage particles are lysed by proteinase K and the protein components of the capsid are degraded. The DNA is purified by phenolic extraction or anion exchange chromatography (Section 26.1). Phage DNA is a high molecular linear DNA (45–50 kB) that should be handled with care. For many applications, phage DNA can also be obtained with commercial kits and the needed fragments amplified by PCR.

26.4.2 Isolation of Eukaryotic Viral DNA

The diversity of eukaryotic viruses requires several additional adapted strategies for isolation of their nucleic acids. Two general purification methods can be distinguished.

In infected cells the viral DNA is present as extrachromosomal DNA (e.g., adenoviruses, polyomaviruses, SV40, papillomaviruses, baculoviruses). The viral DNA can be isolated from the infected cells via Hirt extraction (Chapter 26.3.2) in high yield and sufficient purity for many applications. Because some viruses contain high molecular weight DNA the same precautions should be taken as with any other high molecular weight DNA. The Hirt extraction method does not yield highly pure virus DNA and the viral DNA might be modified differently than in the virus particle (proteins bound, covalent modifications, circular or non-covalently closed DNA).

Highly pure, native viral DNA can be obtained by purification of the viral particles. In most cases, infected cell release newly synthesized virus particles into the medium. The viral particles can be pelleted by ultracentrifugation (approx. 100 000*g*) and purified using CsCl gradient centrifugation. The viral shell is lysed specifically depending on the virus type. Usually, viruses are incubated with proteinase K followed by phenolic extraction. Using this method, proteins bound to viral DNA, like the terminal protein bound to adenoviral DNA or chromatin-like structures of Polyoma or SV40 nucleic acids, are destroyed. In some cases, mild alkaline lysis is sufficient to isolate native viral DNA.

Commercial kits use silica-membrane based matrices or anion exchange chromatography. The viral DNA is isolated using cell free liquids (supernatants of blood plasma) as the methods do not allow the separation of viral and cellular DNA. Using spin columns or 96-well filter plates blood or samples can be analyzed by (RT-)PCR for the presence of viral DNA (or RNA), for example, of HBV, HCV, and HIV on a high throughput basis.

26.5 Isolation of Single-Stranded DNA

26.5.1 Isolation of M13 Phage DNA

Filamentous phages like M13, f1, or fd possess single-stranded covalently closed circular DNA (approx. 6.5 kb). Cloning of foreign DNA into the phage genome allows the isolation of single-stranded DNA of the desired sequence with high yield. The phage M13 infects exclusively *E. coli* (e.g., JM109, JM197) by intrusion using the sex pili of the bacteria that are coded on the F episome. The phage is converted into the replicative form (RF), a double-stranded version of the phage. Infection of the bacteria with M13 does not result in lysis as with bacteriophage λ infections but only in diminished growth rates. The single-stranded version of the M13 genome is isolated by isolation of the phage particles; the replicative form can be purified from the bacterial pellet. M13 phages are isolated by poly(ethylene glycol) precipitation or by anion exchange chromatography.

Commercial kits are available for the isolation of M13 to allow high throughput isolation. The purification is based on silica-gel membranes. At high salt conditions, single-stranded DNA binds with higher affinity to this material than do double-stranded DNA or proteins.

26.5.2 Separation of Single- and Double-Stranded DNA

Single-stranded and double-stranded DNA can be separated from a complex mixture by hydroxyapatite chromatography. Hydroxyapatite, a crystalline form of calcium phosphate ($Ca_5(PO_4)_3(OH)$), is bound preferentially by double-stranded DNA and with much lower affinity by single-stranded DNA or RNA. Binding of double-stranded DNA is performed using phosphate containing buffer at high temperatures (60 °C). At these conditions the single-stranded DNA does not bind to the column and will be found in the void volume. The double-stranded DNA can be eluted by increasing the phosphate content of the buffer.

A critical factor of this purification method is the high phosphate content of the obtained nucleic acids as this interferes with nucleic acid precipitation. The fractionated nucleic acids are concentrated first with *sec*-butanol and then desalted by gel filtration.

26.6 Isolation of RNA

Working with RNA requires even more care than working with DNA. RNases are in contrast to DNases very stable, do not need any co-factors, and cannot be inactivated completely by autoclaving. Only ultrapure buffers should be used for isolation of RNA. RNases can be inactivated by treating the buffers with diethyl pyrocarbonate (DEPC) (Figure 26.10). DEPC inactivates RNAses by covalent modification of the histidine residue in the active center of the enzyme. Buffers containing free amino groups cannot be treated with DEPC. DEPC is toxic due to its modifying properties. Excess DEPC needs to be inactivated by autoclaving as DEPC present during RNA purification will modify the bases of the RNA (carboxyethylation of the adenines and, seldom, guanines). DEPC is degraded to carbon dioxide and ethanol. The use of gloves and sterile, RNase free plastic ware is essential for handling RNA. Glass ware should be decontaminated by heat treatment at 300 °C. For many experiments, RNase inhibitors can be added but these inhibitors can only inactivate low contents of RNases (Table 26.6). In addition

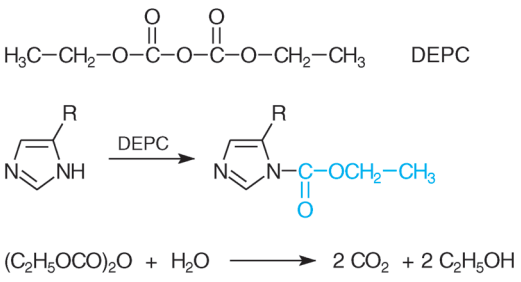


Figure 26.10 Chemical formula and mechanism of DEPC (diethyl pyrocarbonate) treatment. DEPC inactivates RNases by covalent modification of amino groups and histidines. DEPC degrades to ethanol and carbon dioxide upon heating and autoclaving.

Table 26.6 Frequently used RNase inhibitors.

RNase inhibitor	
RNasin	Protein from human placenta Forms non-covalent equimolar complexes with RNases Cannot be used at denaturing conditions
Diethyl pyrocarbonate	Used for buffer treatment Covalent modification Needs to be inactivated
Vanadyl ribonucleoside complexes	Transition state analog that binds RNases and inhibits their activity Cannot be used for cell free translation systems
SDS, sodium deoxycholate	Denaturation
β-Mercaptoethanol	Reduction
Guanidinium thiocyanate	Used in connection with cell lysis Denatures RNases reversibly
Formaldehyde	Used in denaturing agarose gels Covalent modification

to these commonly used RNase inhibitors, optimized protein- or antibody-based inhibitors specific for certain RNases are commercially available.

26.6.1 Isolation of Cytoplasmic RNA

In contrast to DNA localized in the nucleus, most RNA molecules are located in the cytoplasm. Cytoplasmic RNA is composed of various RNA species, such as classical, long-known ribosomal RNA, transfer RNA, and messenger RNA. With recent new technologies like deep sequencing and tiling arrays, new RNA species have been identified. A majority of the human genome is transcribed while only an estimated 2% of these RNAs will be translated to proteins. The so-called non-coding RNAs constitute a new group of RNA molecules with various functions, many of which have yet to be discovered. Non-coding RNAs with sizes above 200 nt are classified as long ncRNAs, whereas miRNA (micro), piRNA (PIWI-interacting), and siRNA (small interfering) belong to the group of small ncRNAs.

Due to the very heterogeneous nature of the RNAs, various isolation and purification protocols and commercial kits are available. For some applications, like Northern blots or RT-PCR or ribonuclease protection assays, the isolation of cytoplasmic RNA is sufficient. Minor contamination of the RNA preparation with genomic DNA can be excluded by the use of proper controls. A RT-PCR reaction, for example, can be performed without the use of reverse transcriptase. The PCR result should be negative if no genomic DNA is present. In the presence of contaminating genomic DNA the PCR will be positive even without reverse transcriptase treatment. The use of intron/exon spanning primers is recommended. With these primer pairs, only spliced mRNA will yield fragments of the correct size. For some applications it can be useful to enrich or purify the mRNA out of the total RNA.

Cultivated Cells The plasma membranes of the cells are lysed with a non-ionic detergent (Nonidet P40) while keeping the cell nuclei intact. The nuclei are separated and the proteins in the cytoplasmic fraction are degraded using proteinase K. The RNA can be purified by phenolic extraction. If cells have been transfected with plasmid DNA, the cytoplasmic RNA can be contaminated with episomal DNA, which can be removed by digestion with RNase-free DNases.

Tissue and Cultivated Cells The nuclease activity of a tissue can be very high. Therefore, the tissue is frozen immediately in liquid nitrogen. Cells are lysed and proteins completely denatured using the chaotropic salt guanidinium thiocyanate. β-Mercaptoethanol and L-lauryl-sarcosine (Figure 26.11) are added to prevent degradation of the RNA. Cells or tissue are also often lysed using phenol. Since RNases are not completely inactivated by pure phenol, a mixture of acidic phenol: chloroform: isoamyl alcohol (Section 26.1) is used. Most methods and kits use the combination of both reagents for efficient and more convenient denaturation of proteins and inactivation of RNases.

The RNA can be purified using anion exchange in a similar manner to DNA purification. For RNA, adapted buffer conditions are used to bind and elute the RNA from the column.

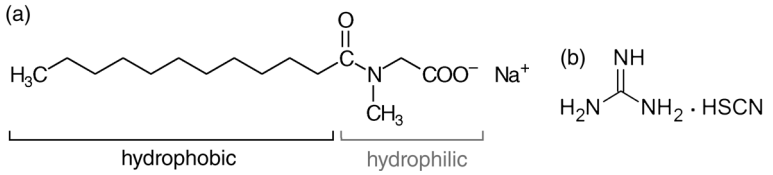


Figure 26.11 (a) *N*-Lauryl-sarcosine and (b) guanidinium thiocyanate.

Contaminating DNA can be removed by digestion (also on-column) with DNases. RNA can also be purified using CsCl density gradient centrifugation. RNA–ethidium bromide complexes pellet due to their higher density and can so be separated from genomic DNA. If the RNA needs to be purified as a band, higher density gradients (using Cs_2SO_4) need to be performed (Section 26.3.1).

Most commercial kits are based on silica technology use solid phase extraction (SPE). RNA (and also DNA) can be bound to filters or columns consisting of silica particles, glass fibers, or glass beads (Section 26.7). The RNA is bound to the silica-material in the presence of high-salt chaotropic buffers (in most cases this buffer is provided during lysis of cells using guanidinium thiocyanate buffer). The RNA is washed and eluted from the matrix with low salt buffers; in many commercially available kits, the eluent is RNase free water. These recent technologies enable researchers to obtain high quality RNA simply and quickly in high throughput quantities. All RNA isolation kit sellers offer specialized protocols and kits for all kind of RNA sources and applications.

For some applications, for example, next generation sequencing, it is useful to remove the major part of the ribosomal RNA before sequencing. This will reduce material cost due to unnecessary sequencing of contaminating ribosomal RNA. Ribosomal RNA depletion kits are based on hybridization of the ribosomal RNA to oligonucleotide probes specific for ribosomal RNA. The hybridized ribosomal RNA:DNA strands are then bound to beads and removed from the solution, for example, by magnetic separation (Section 26.7).

26.6.2 Isolation of Poly(A) RNA

Nearly all eukaryotic mRNA species contain long adenine-rich regions on their 3′ termini. These poly(A) tails are used to purify mRNA from cytoplasmic RNA. Column or bead material are coupled to single-stranded thymidine rich short DNA fragments (oligo(dT)). The poly(A) tails hybridize to the oligo(dT) strands and are bound to the column material (Figure 26.12). Contaminating non-poly(A) containing RNAs can be easily removed by washing the column.

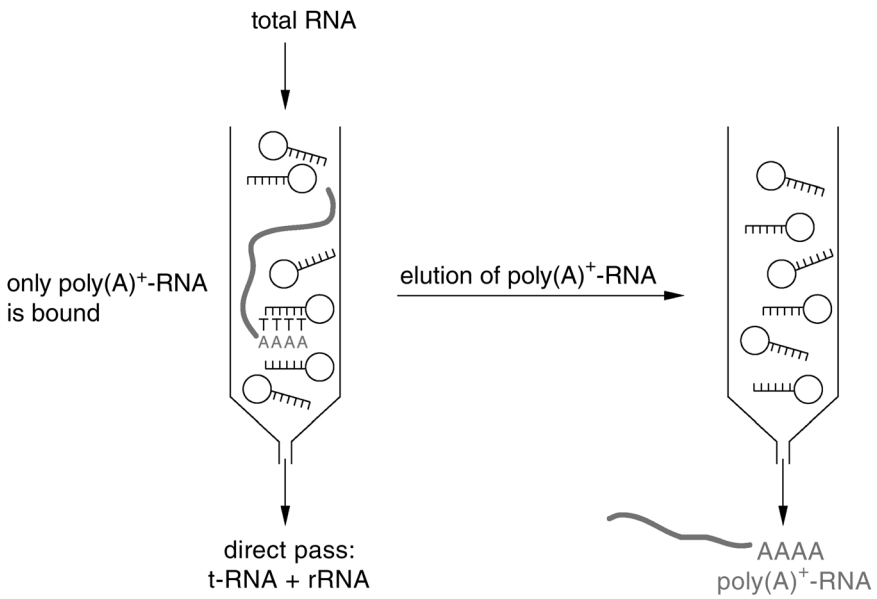


Figure 26.12 Isolation of poly(A) RNA using a oligo(dT) column. Total RNA (cytoplasmic RNA) is loaded to the column. RNA with a poly(A) tail is bound by hybridization of the adenines to the oligo (dT) residues to the column whereas all other molecules are collected in the flow through. The poly(A) RNA is eluted using conditions that destabilize the dT:rA hybrids.

To ensure optimal hybridization and loading of the column, the starting RNA material needs to be denatured. For optimal yield, the starting material can be applied to the column several times. The poly(A)-RNA is bound to the column at high salt concentrations (500 mM NaCl or LiCl) and purified poly(A) RNA is eluted with water. These conditions destabilize the dT:rA hybrids. Low cost oligo(dT) columns can be prepared by coupling of oligonucleotides (dT12–18) to activated column material. For more convenience, commercial kits are available in different formats.

26.6.3 Isolation of Small RNA

In recent years significant research has focused on small non-coding RNAs like miRNAs, siRNAs, or piRNAs with sizes lower than 200 nt. These RNAs are purified from tissues, cells, or extracellular vesicles like exosomes.

Many of the RNA isolation and purification protocols developed for longer RNAs are not optimal for small RNAs, for example, ethanol precipitation is not very efficient for small RNAs and many protocols need to be adapted. It is important for a good recovery of the small RNAs to include an acid phenolic extraction at the beginning of the isolation protocol. Only if tissue or cells or exosomes are denatured completely using acidic phenol : chloroform : isoamyl alcohol is the yield of small RNAs sufficient. Individual purification protocols depend on the column material and kit used and specific enrichment of small RNAs can be achieved by a combination of different separation techniques and buffer conditions. Tailored isolation kits are accessible to the research community for the purification of small RNAs from different sources

26.7 Isolation of Nucleic Acids using Magnetic Particles

In recent years, the demands on nucleic acid purification protocols have increased dramatically regarding speed, costs, yield, purity, and format. A lot of scientific questions require isolation of a huge number of samples simultaneously, for example, for expression profiling or SNP analysis (single nucleotide polymorphism). The development of automated high throughput isolation protocols was mandatory. Certain protocol steps cannot be transferred easily to automated liquid handling systems (e.g., centrifugation). New protocols needed to be developed. The isolation of nucleic acids can easily be automated using magnetic particles. Beads with paramagnetic (will be magnetized by an external magnetic field) or magnetic properties are used. Applications of this technique are very general and have advantages compared to the conventional separation protocols. The material is not subject to shear forces as no centrifugation steps are necessary and the use of organic reagents is obsolete. The magnetic beads are loaded with the nucleic acids and brought to an external magnetic field. The beads and bound nucleic acids are retained in the magnetic field while decontaminating material can be washed away (Figure 26.13). If used manually, the beads are often retained in a column placed in a magnetic field. In automated liquid handling systems, the magnetic field is usually provided by a magnetic plate, on which the 96-well plate containing the beads is placed.

For the isolation of DNA, silica coated magnetic beads are used as DNA binds in the presence of chaotropic reagents to glass surfaces (Section 26.5). Using solid phase reversible immobilization (SPRI), DNA is loaded reversibly to magnetic beads modified with carboxyl-groups in the

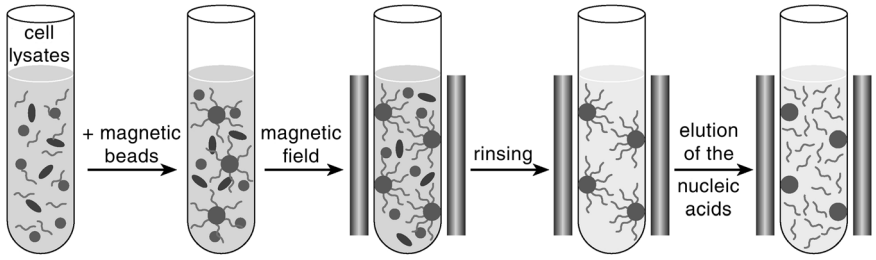


Figure 26.13 Principle of magnetic bead isolation. The nucleic acids in cell or bacterial lysates are bound specifically to the magnetic particles. By applying a magnetic field the beads are fixed and the contaminations can be washed away. After the washing steps the nucleic acids are eluted from the magnetic beads. All protocol steps can be performed on automated systems. The isolation protocol and the kind of bead depend on the type of nucleic acid to be purified.

presence of high salt concentration and poly(ethylene glycol) (PEG). The PEG is important for the binding of the DNA to the bead surface. Streptavidin coated beads are used for the isolation of very low amounts of mRNA. The beads are coupled with biotinylated oligo(dT) primer and the mRNA is coupled and isolated. This principle of binding biotinylated nucleic acids to streptavidin beads can be applied to a huge variety of isolation methods (e.g., isolation of DNA binding proteins).

26.8 Lab-on-a-chip

Not only the format, time, and throughput of nucleic acid purification protocols have improved significantly in recent decades – with the lab-on-a-chip (LOC) system it is possible to isolate DNA in a miniaturized fashion. The LOC is part of the microelectromechanical systems (MEMSs) and is based on a chip that is between square millimeters and centimeters in size. The volume can be as low as 1 picoliter. The concept is to include all techniques, starting from the isolation of the nucleic acids (from blood or tissue) to the analysis of the nucleic acids, on the same chip. The systems are also part of the micro total analysis systems, μTAS. Similar to the isolation methods for automated liquid handling systems, protocols cannot be based on centrifugation or phenolic extraction. It is also important to achieve a certain concentration of the DNA for the subsequent analysis steps. The SPE methods can be transferred in part to the chip technology. The silica based isolation methods where DNA binds to the solid phase in the presence of chaotropic reagents are suitable. In addition, SPRI methods are applied. Additional suitable materials, like poly(methyl methacrylate) (PMMA), are used to enlarge the active surface on the chip.

Further Reading

Ausubel, E.M., Brent, R., Kingston, R.E., Moore, D.D., Smith, J.A., Seidman, J.G., and Struhl, K. (1987) *Current Protocols in Molecular Biology*, John Wiley & Sons, Inc., New York.

Farrell, R.E. (2010) *RNA Methodologies: A Laboratory Guide for Isolation and Characterization*, 4th edn, Academic Press, Elsevier.

Glasel, J.A. and Deutscher, M.E. (1995) *Introduction to Biophysical Methods for Protein and Nucleic Acid Research*, Academic Press, New York.

Green, M.R. and Sambrook, J. (2012) *Molecular Cloning: A Laboratory Manual*. 4th edn, Cold Spring Harbour Press, Cold Spring Harbor.

Krieg, P.A. (ed). (1996) *A Laboratory Guide to RNA: Isolation, Analysis and Synthesis*, Wiley-Liss, New York.

Kües, U. and Stahl, U. (1989) Replication of plasmids in Gram-negative bacteria. *Microbiol. Rev.*, **53**, 491–516.

Levinson, P., Badger, S., Dennis, J., Hathi, P., Davies, M., Bruce, I., and Schimkat, D. (1995) Recent developments of magnetic beads for use in nucleic acid purification. *J. Chromatogr. A*, **816**, 107–111.

Micklos, D.A. and Freyer, G.A. (1990) *DNA Science. A First Course in Recombinant DNA Technology*, Cold Spring Harbor Laboratory Press and Carolina Biological Supply Company, Cold Spring Harbor.

Perbal, B. (1998) *A Practical Guide to Molecular Cloning*, John Wiley & Sons, Inc., New York.

Price, C.W., Leslie, D.C., and Landers, J.P. (2009) Nucleic acid extraction techniques and application to the microchip. *Lab Chip*, **9**, 2484–2494.

Tan, S.C. and Yiap, B.C. (2009) DNA, RNA and protein extraction: the past and the present. *J. Biomed. Biotechnol.*, article ID 574396.

Wen, J., Legendre, L.A., Bienvenue, J.M., and Landers, J.P. (2008) Purification of nucleic acids in microfluidic devices. *Anal. Chem.*, **80**, 6472–6479.

Zähringer, H. (2012) Old and new ways to RNA. *LabTimes*, (2), 52–61.

Analysis of Nucleic Acids

27

Nucleic acids, isolated from different sources, different tissues from different organisms or cell or tissue cultures, subsequently appear as a compact, high molecular bulk of, especially in case of genomic DNA, unspecific fragments, which are hard to analyze in this status. For processing it is necessary to determine purity, conformation, fragment size and last but not least the sequence of these nucleic acid fragments.

In this chapter we summarize basic analytical methods available for nucleic acids processing. The presented methods result in a basic characterization and/or are necessary for more detailed and final characterization or manipulation of nucleic acids. For example the transformation of a high molecular bulk of nucleic acids into specific molecular fragments by restriction analysis, which can easily be further characterized and manipulated for example by cloning. Fragments can be separated by gel electrophoresis, visualized by staining, isolated from the gel matrix or transferred by "blotting" to a specific carrier material for more specific characterization by "hybridization". Most of these techniques are basic and daily routines when working with nucleic acids.

27.1 Restriction Analysis

Ute Wirkner[1] and Joachim W. Engels[2]

[1] German Cancer Research Center, Clinical Cooperation Unit Translational Radiation Oncology, Im Neuenheimer Feld 400, 69120 Heidelberg, Germany
[2] Goethe University Frankfurt, Institute of Organic Chemistry and Chemical Biology, Department of Biochemistry, Chemistry and Pharmacy, Max-von-Laue Straße 7, 60438 Frankfurt am Main, Germany

Restriction analysis is used for the characterization, identification, and isolation of double-stranded nucleic acids and, thus, is a basic tool in nucleic acid analysis. Cloning of DNA molecules is almost unthinkable without restriction analysis. Even if it is possible to clone DNA by PCR without the need of restriction, mostly restriction analysis is used to prepare the DNA fragments and the vectors for cloning and to identify the resulting cloning product. In addition, for any other kind of DNA manipulation, such as mutagenesis or amplification by PCR, restriction analysis is the tool of choice to identify the desired product. To initially determine the crude structure of any DNA, from small fragments to whole genomes, establishing a restriction map is a useful step on the way to complete sequencing. Restriction analysis of genomic DNA to detect mutations or restriction fragment length polymorphisms (RFLPs) is used for genetic mapping, to identify and isolate disease genes, or, for example, in criminalistics to identify individuals.

27.1.1 Principle of Restriction Analyses

The basis for restriction analysis is the activity of restriction enzymes, which bind and cut double-stranded DNA molecules at specific recognition sequences. These are mainly so-called

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

type II restriction enzymes. DNA-fragments resulting from this activity have a specific length, defined by the positions of the recognition sites, and can be separated according their size by gel electrophoresis (Section 27.2.1). The analysis of a DNA molecule results in a specific band pattern of restriction fragments. By comparison with a respective size standard each fragment can be assigned to its approximate size. Depending on the size of the initial DNA molecule(s) the *detection* of the fragments is carried out: if the original molecule is comparatively small, as with most cloning products or vectors (plasmids, lambda phages, cosmids around 3–50 kb), unspecific detection by staining all nucleic acids in the gel by, for example, ethidium bromide is sufficient (Section 27.3.1). If a certain area within a complex genome is to be analyzed, detection by specific Southern blot hybridization (Section 27.4.3) has to be performed or the fragment has to be amplified *in vitro* by PCR before undergoing a restriction analysis. Thus restriction analysis can be performed on any type and size of double-stranded DNA and is comparatively easy and quick to perform. Another undoubted aspect of the broad application spectrum is the large variety of restriction enzymes and respective restriction sites.

Hybridization Section 28.1
PCR, Chapter 29

27.1.2 Historical Overview

Before restriction enzymes were discovered it was almost impossible to characterize or isolate certain genes or genomic areas. Genomic DNA isolated from cells or tissues is a mass of large, chemically monotonous molecules that in principle only allow separation according to their size. Since a functional unit such as a gene does not exist as a single molecule in a cell but is part of a much larger DNA- molecule, specific breakdown of this large molecule is required to separate and isolate an interesting section (e.g., a gene). Even if DNA–molecules can be sheared at random sites by mechanical forces the result is again a heterogeneous unspecific mixture of DNA molecules and no defined DNA-fragments can be isolated. It was only with the discovery and isolation of restriction enzymes by Arber, Smith, and Nathans in the late 1960s that the opportunity first arose to manipulate DNA in a defined way, namely, to degrade it to specific fragments of defined length, which can be specifically separated and isolated. Hence, a first detailed characterization of DNA was possible and the basis for isolating and amplifying DNA by cloning was laid. By the implementation of restriction enzymes and consecutive cloning, hybridization, and other enzymatic DNA-manipulations, DNA has shifted from the least accessible to the easiest manipulable and analyzable macromolecule.

27.1.3 Restriction Enzymes

Restriction enzymes are endonucleases that occur mainly in bacteria but they also exist in viruses and eukaryotes. They cleave phosphodiester bonds of both strands in a DNA molecule by hydrolysis and differ in recognition sequence, cleavage site, organism of origin, and structure. Several thousand restriction enzymes with several hundred different recognition sequences are known.

Biological Function The biological function of restriction enzymes is to protect the organism of origin from infiltration by foreign DNA (e.g., from phages by cleaving and inactivating and, thus, "restricting" the growth of the phage). Its own DNA and every DNA synthesized in the cell is protected from this attack by modifications, mostly methylation. This restriction/modification (R/M) system is specific to its organism of origin and is a protection mechanism, a kind of immune system. The host specificity of bacteriophages is based on this system; they only can infect bacteria efficiently that have the same methylation pattern as their bacteria of origin.

Classification of Restriction Enzymes Primarily, three types of restriction enzymes (I, II, and III) are distinguished, whose properties and differences are summarized in Table 27.1. Type I restriction enzymes possess restriction as well as methylation activity and have a defined recognition sequence. If both strands of the recognition sequence are methylated, DNA will not be cleaved. If only one strand is methylated, the sequence is recognized and the second strand will be methylated. If both strands are methylated the sequence is recognized as well and the DNA is cleaved around 100 bp away from the recognition sequence. Restriction enzymes that are most

Table 27.1 Classification of restriction enzymes (REases).

	Type I	Type II	Type III
Function	Endonuclease and methylase	Endonuclease	Endonuclease and methylase
Recognition sequence	Two parts, asymmetric	4–8 Bases, most palindromic	5–7 Bases, asymmetric
Cleavage site	Unspecific, often >1000 bp distance to the recognition sequence	Within or close to recognition sequence	5–20 Bases in front of the recognition sequence
ATP needed	Yes	No	Yes

frequently used in analytics are usually type II restriction enzymes. In addition, most well-known restriction enzymes belong to type II. In contrast to type I and type III restriction enzymes, type II restriction enzymes usually possess only restriction activity and cleave the DNA within the recognition sequence, which results in DNA fragments of defined length and defined ends. Type III restriction enzymes, like type I, have restriction activity as well as methylation activity. They cleave the DNA at sites at a distinct distance from the recognition sequence so that the resulting fragments have a defined length and variable ends.

So-called homing endonucleases, like I-PpoI, in their native form have longer recognition sequences (>15 bp) and initiate the insertion of their own genes, so-called homing endonucleases genes (HEGs). They are selfish elements that colonize genomes and occur in different animal kingdoms from bacteria to eukaryotes. These endonucleases are mainly used for gene targeting and they are engineered to alter target site specificity. Nickases represent a small group of nucleases that have double-stranded recognition sequences but cleave only one strand of the DNA.

Nomenclature of Type II Restriction Enzymes (REases) The nomenclature of type II restriction enzymes is based on the organism of origin. For example, restriction enzyme EcoRI was isolated from a resistance (R) factor of *Escherichia coli* strain RY 13. Here "I" stands for the first restriction enzyme isolated from this strain. Analogous BamHI was the first enzyme isolated from *Bacillus amyloliquefaciens* strain H. The scientific community agreed in this unique nomenclature in 2003. Here the terms restriction enzyme and restriction endonuclease were denoted synonymously and the abbreviation REases was introduced. Since type II enzymes compose by far the largest group of restriction enzymes, and since in addition there are members that differ from classical recognition features, the type II group was segmented into subgroups, which are described below in Table 27.3. All type II enzymes do not depend on ATP, they mostly do not form a complex with the respective methylase, they recognize a specific DNA sequence, and they cut within or close to the recognition sequence. The resulting DNA fragments have 5′-phosphate and 3′-OH groups.

Recognition Sequences Thousands of type II restriction enzymes with hundreds of recognition sequences have been characterized, and the number is constantly increasing. Comprehensive compilations can be found in regularly updated databases, company catalogues, or books of molecular biological methods. The recognition sequences of these restriction enzymes span 4–8 nucleotides and are, most often, palindromic. Table 27.2 lists representative examples of some restriction enzymes and their recognition sequences. In accordance with convention the sequence is given in the direction 5′ to 3′.

The cleavage site is usually located within the recognition sequence and thereby the resulting restriction fragments have defined ends, which is among other factors relevant for cloning. But there are also restriction enzymes like FokI (Table 27.3), whose cleavage site is a few bases away from the recognition site. As shown in Figure 27.1, the cleavage of DNA with restriction enzymes can result in blunt ends or in cohesive or "sticky" ends. Sticky ends can have either an overhanging 5′ or 3′ end, depending on which strand of the DNA forms the overhang. Usually, DNA fragments resulting from restriction enzyme activity have a 3′ hydroxyl and a 5′ phosphate group.

Table 27.2 Specification of some type-II restriction enzymes (type-II-REases).

Restriction enzyme	Recognition and cleavage site[a)]	Organisme of origin	Isoschizomers
*Bam*HI	G/GATCC	*Bacillus amyloliquefaciens* H	*Bst*I
*Bst*I	G/GATCC	*Bacillus stearothermophilus* 1503-4R	*Bam*HI
*Eco*RI	G/ÁATTC	*Escherichia coli* RY13	
*Fok*I	$GGATGN_9$/ $CCTACN_{13}$/	*Flavobacterium okeanokoltes*	
*Hind*II	GTPy/PuAC	*Haemophilus influenzae* R_d	*Hind*II
*Hind*III	A/AGCTT	*Haemophilus influenzae* R_d	
*Hpa*II	C/CGG	*Haemophilus parainfluenzae*	*Msp*I
*Msp*I	C/CGG	*Moraxella* species	*Hpa*I
*Not*I	GC/GGCCGC	*Nocardia otitidiscaviarum*	
*Sad*I	GAGCITC	*Streptomyces achromogenes*	
*Sau*3A	/GATC	*Staphylococcus aureus* 3A	*Mbo*I, *Nde*II
*Sma*I	CCC/GGG	*Serratia marcescens* Sb	*Xma*I
*Xma*I	C/CCGGG	*Xanthomonas malvacearum*	*Sma*I

a) Py: pyrimidine (C or T); Pu: purine (A or G); N: A, C, G, or T.

The frequency of a recognition sequence depends mainly on its length but also on its own base composition and the composition of the DNA that is restricted. Assuming a random composition a 4 bp recognition sequence statistically occurs approximately every 4^4 bp (256 bp), a 6 bp or 8 bp recognition sequence respectively every 4^6 bp (4096 bp) or 4^8 bp (65 536 bp). However, different organisms possess different base compositions of their

Table 27.3 Subtypes of Type-II restriction enzymes (REases).

Subtype[a)]	Characteristics	Example	Recognition and cleavage site
A	Asymmetric recognition sequence	*Fok*I *Acr*I	GGATG (9/13) CCGC (3/-1)
B	Cleavage on both sides of the recognition sequence	*Bcg*I	(10/12) CGANNNNNNTGC (12/10)
C	Symmetric or asymmetric recognition sequence; R- and M-function in one polypeptide	*Gsu*I *HaeIV* *Bcg*I	CTGGAG (16/14) (7/13) GAYNNNNNRTZ (14/9) (10/12) CGANNNNNNTGC(12/10)
E	Two copies of the recognition sequence, one is cleaved, the other serves as an allosteric effector	*Eco*RII *Nae*I	↓CCWGC GCC↓GGC
F	Two recognition sequences, both are cleaved in coordination	*Sfi*I *Sgr*AI	GGCCNNNN↓NGGCC CR↓CCGGYG
G	Symmetric or asymmetric recognition sequence; depend on AdoMet	*Bcg*I *Eco*571	GTGCAG (16/14) CTGAAG (16/14)
H	Symmetric or asymmetric recognition sequence, gene structure similar to type I REases	*Bcg*I *Ahd*I	(10/12) CGANNNNNNTGC(12/10) GACNNN↓NNGTC
M	Subtype IIP or IIA; recognize only methylated recognition sequences	*Dpn*I	Gm6A ↓TC
P	Symmetric recognition and cleavage site	*Eco*RI *Ppu*MI *Bsl*I	G↓AATTC RG ↓GWCCY CCNNNNN↓NNGG
S	Asymmetric recognition and cleavage site	*Fok*I *Mme*I	GGATG (9/13) TCCRAC (20/18)
T	Symmetric or asymmetric recognition sequences; heterodimers	*Bpu*10I *Bsl*I	CCTNAGC (–5/-2)[b)] CCNNNNN↓NNGG

a) Not all subtypes are exclusive! For example, *Bsl*I is subtype P and T.
b) Abbreviation means the following cleavage: 5′ CC ↓T N AG C 3′ G GA N T ↓C G

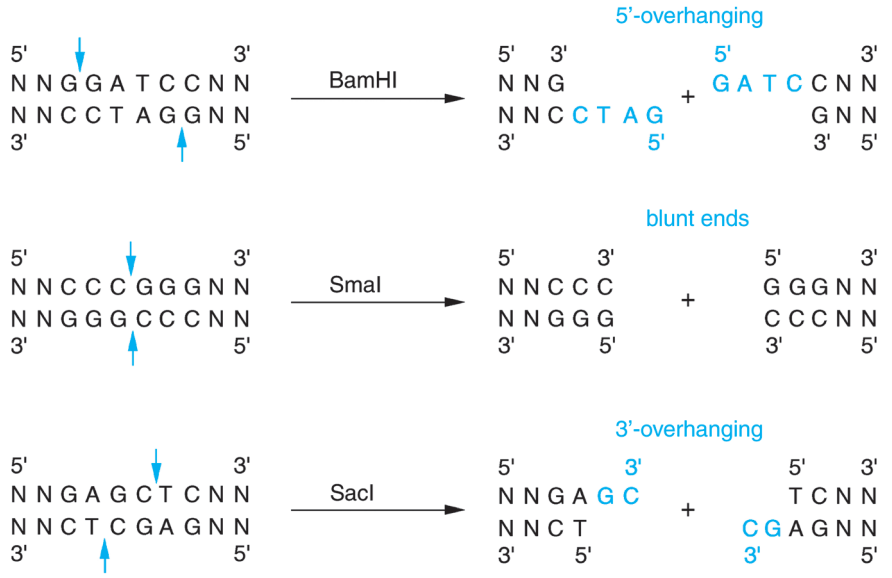


Figure 27.1 DNA ends generated by restriction enzyme cleavage. Depending on the applied restriction enzyme three kinds of DNA ends occur: cohesive ends (*sticky ends*) arise by, for example, cleavage with BamHI and SacI, whereby BamHI creates 5′-overhanging and SacI 3′-overhanging ends. Blunt ends are created by, for example, SmaI.

genomes The A/T and accordingly the G/C content is rarely 50% and the dinucleotide CpG occurs less frequently in eukaryotes than do the other dinucleotides. Consequently, a recognition sequence containing CpG will occur less frequently in eukaryotic genomes than calculated according its length. Restriction enzymes with an 8 bp recognition sequence are, for example, applied to establish restriction maps from whole chromosomes. The resulting, very long DNA fragments are separated and detected by pulse-field gel electrophoresis (Section 27.2.3). Most frequently used restriction enzymes recognize 6 bp sequences since the length of the resulting fragments is good for separation and isolation. However, if a partial restriction is to be performed, for instance to establish a genomic library, restriction enzymes with 4 bp recognition sequences are selected.

Isoschizomeres Isoschizomeres (Table 27.2) are restriction enzymes that have identical recognition sequences but originate from different organisms. The cleavage site might be identical (e.g., *Bam*HI and *Bst*l) or different (e.g., *Sma*l and *Xma*I). Isoschizomers with different cleavage sites are termed neoschizomers. The enzymes might also differ in their sensitivity towards methylation: for example, *Hpa*II and *Msp*l have identical recognition sites but *Hpa*II does not cleave if the second cytosine is modified to 5-methylcytosine (^{5m}C), while *Msp*l will cleave despite this methylation.

27.1.4 *In Vitro* Restriction and Applications

In a restriction enzyme reaction mixture, the DNA to be analyzed is incubated with the desired restriction enzyme under defined buffer conditions at a defined temperature for a certain time. The restriction buffer usually contains Tris-buffer, $MgCl_2$, NaCI, or KCl as well as a sulfhydryl reagent (dithiothreitol (DTT), dithioerythritol (DTE) or 2-mercaptoethanol). A divalent cation (mostly Mg^{2+}) is necessary for enzymatic activity, as well as the buffer, which provides the correct pH, mostly between pH 7.5 and pH 8. Some restriction enzymes are sensitive towards ions such as Na^+ or K^+, while others are active within a wide concentration range. Sulfhydryl reagents stabilize the enzyme. The optimal temperature for most restriction enzymes is 37 °C, but it may vary depending on the enzyme, with respect to the organism of origin, to higher (e.g., 65 °C for *Taq*l) or lower (e.g., 25 °C for *Sma*l) temperatures.

The amount of restriction enzyme is given in units: one unit of a restriction enzyme is the amount needed to cleave one microgram of substrate DNA under optimal conditions within one hour. As a general rule bacteriophage lambda DNA is used as substrate for this definition.

Complete Restriction For most purposes complete restriction of DNA is intended. For this, optimal conditions for the respective restriction enzyme are selected and a sufficient amount of enzyme for the DNA to be cleaved.

Incomplete or Partial Restriction For some purposes, like restriction mapping or the preparation of a genomic DNA-library, partial restriction is desired. This means that statistically not all of the restriction sites are cleaved. This is achieved by an under optimization of the reaction conditions, such as a lower amount of restriction enzyme, shorter reaction time, or change of buffer conditions (e.g., reduced $MgCl_2$ concentration).

Multiple Restriction This involves the restriction of DNA with several restriction enzymes. The DNA might be incubated either simultaneously or one after another with the desired restriction enzymes. The crucial criterion here is the compatibility of the reaction conditions. Multiple restriction, among others methods, is applied to establish restriction maps.

Restriction Mapping To establish a restriction map recognition sequences of one or several restriction enzymes are localized within a DNA-molecule. Thus the restriction map is a crude physical map of the DNA-molecule to be analyzed. The perfect physical map is the complete nucleotide sequence of the DNA. Consequently, restriction mapping is applied to identify known sequences, for example, to verify successful cloning of a known DNA-fragment or as the first step of projects that aim to identify a complete nucleotide sequence. Therefore, restriction analysis of DNA fragments integrated in cloning vectors (e.g., plasmids, cosmids, or lambda phages) is performed. Before the introduction of next generation sequencing, DNA had to be cloned before elucidating the nucleotide sequence, which was usually performed by Sanger sequencing. Restriction maps of these sequencing clones are established, overlapping clones are identified by comparing their restriction maps, and finally the map of the originally cloned DNA can be elucidated.

Often it is not necessary to isolate the fragments after the first restriction, instead it is sufficient to compare the fragment pattern of the single digest with that of the double digest to determine the order of the restriction fragments. For this approach it is important to use restriction enzymes that produce at least a few overlapping fragments. Consequently, it might be necessary to test some restriction enzymes.

Combination of Multiple Restriction Enzymes By this method the relative position of recognition sequences of different restriction enzymes is determined and from this their absolute position on the originally analyzed DNA fragment. To do so, the DNA fragment to be analyzed is first restricted with each single restriction enzyme in one reaction and fragments are analyzed by gel electrophoresis. Ideally these fragments are isolated and restricted with the second restriction enzyme and these double restrictions again analyzed by gel electrophoresis. By comparing the lengths of the resulting DNA fragments after single and double restriction, overlapping parts can be identified and the relative order of the fragments can be determined. This is shown in Figure 27.2 using the example of a 5 kb DNA-fragment.

Partial Restriction By this method the order of recognition sequences of a single restriction enzyme can be identified. The DNA fragment to be analyzed is digested once completely and once incompletely with the same restriction enzyme and both reactions are analyzed by gel electrophoresis. By comparing the pattern of the resulting restriction fragments the completely restricted fragments can be allocated to the incompletely restricted fragments and thus the order on the original DNA fragment can be determined. This method is shown in Figure 27.3 on a 5 kb DNA-fragment.

In the case of a complex restriction pattern, for example, when analyzing a long DNA fragment or using a very frequently cutting enzyme, it is advisable to use the method shown in Figure 27.4. Hereby, the DNA-molecule is labeled at one end before partial digestion by, for example, incorporation of a labeled nucleotide. After gel electrophoresis these labeled fragments can be detected selectively (e.g., by autoradiography). The size of a detected fragment corresponds here to the distance of a cleavage site to the labeled end of the DNA molecule.

Radioactive Systems, Section 28.4.2

Restriction Analysis of Genomic DNA When carrying out restriction analysis of large eukaryotic genomes there is the problem that too many restriction fragments are generated. After gel electrophoresis no single bands are visible but instead there is a smear of DNA, which consists of specific DNA fragments with many sizes. By selecting certain hybridization probes, a fragment in the analyzed genome containing DNA complementary to the selected probe can be detected. This is done by so-called Southern blot analysis (Section 28.4.3). This analysis enables, for instance, the restriction analysis of a gene whose transcript has been cloned as cDNA and can be used as hybridization probe. There are other objectives for which restriction analysis is helpful, such as the detection of a methylation pattern that is lost by cloning and

Probes for Nucleic Acid Analysis, Section 28.2

Analysis of Epigenetic Modifications, Chapter 31

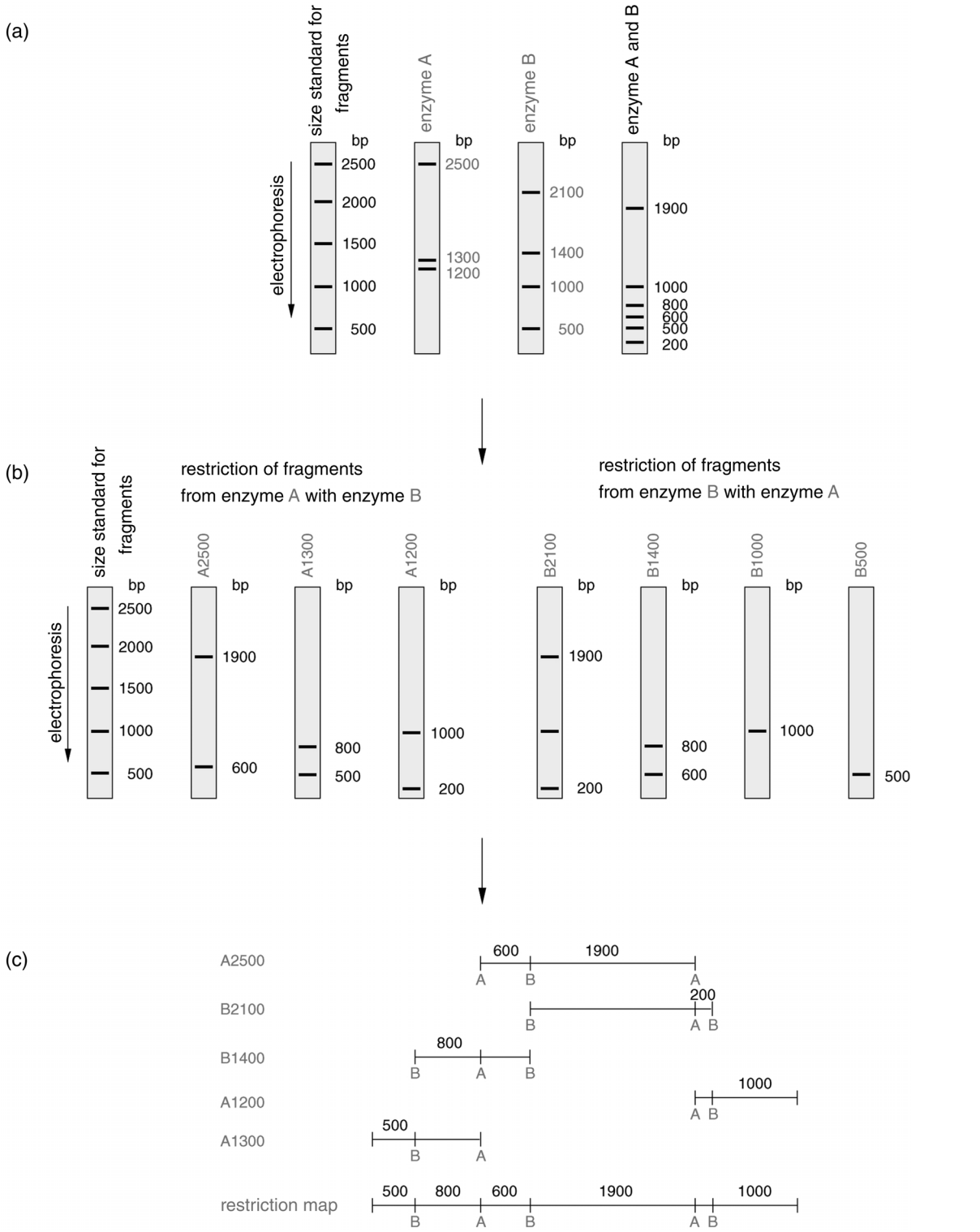


Figure 27.2 Restriction mapping by multiple restriction. A 5 kb long, linear DNA fragment was cleaved by restriction enzymes A and B in single reactions and in a double reaction. (a) Separation of the restriction fragments in an agarose gel. Fragment sizes determined by comparison to the size standard are given. Cleavage with enzyme A results in restriction fragments with lengths of 2500 bp (fragment A2500), 1300 bp (A1300), and 1200 bp (A1200). The corresponding nomenclature for enzyme B fragments and double restriction fragments is shown. Restriction fragments from single reactions were isolated and cleaved with the respective second restriction enzyme: A-fragments with enzyme B and B-fragments with enzyme A. (b) Electrophoretic separation of these secondary cleavage products. By comparison of the restriction pattern, overlapping fragments can be identified and, as shown in (c), can be aligned: The 1900 bp fragment from the double digest is contained in A2500 and B2100; consequently, A2500 and B2100 overlap in this area. In addition, A2500 contains a 600 bp fragments that is also present in B1400, and B1400 contains a 200 bp fragment that overlaps with A1200. After analysis of all fragments the restriction map of the 5 kb DNA fragment can be generated.

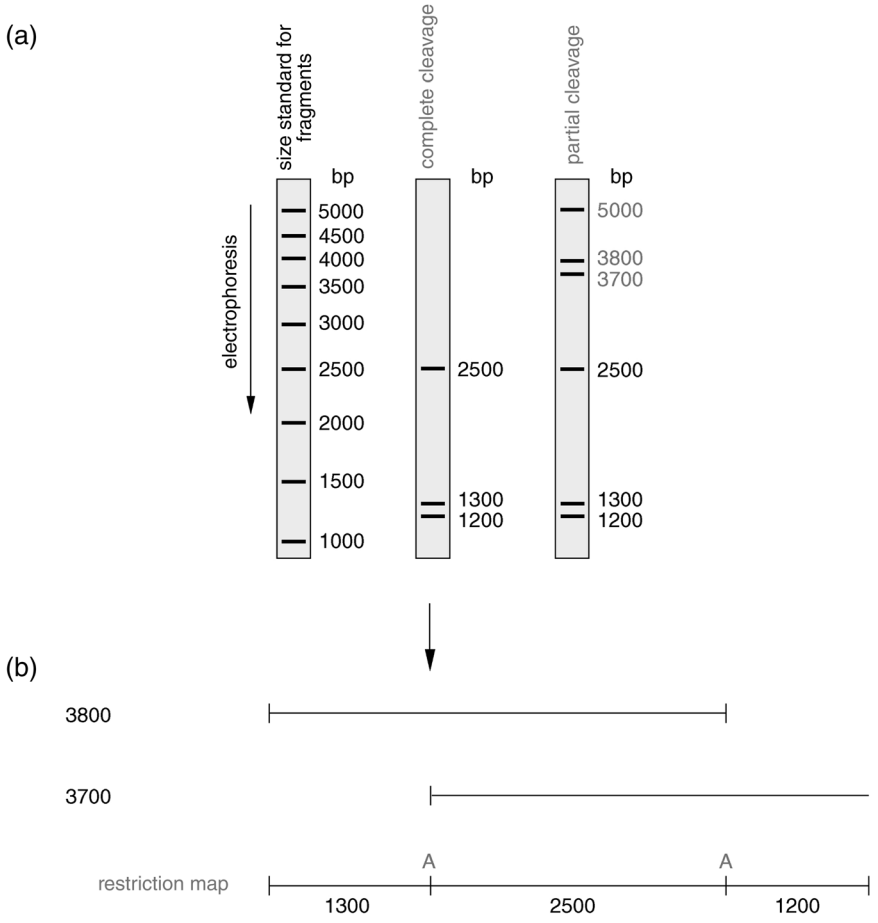


Figure 27.3 Restriction mapping by partial digestion. A 5 kb DNA molecule was cleaved both completely and partially with restriction enzyme A. (a) Gel electrophoretic separation of the resulting restriction fragments. By comparing complete and partial cleavage 5000, 3800, and 3700 bp fragments can be identified as partially cleaved, of which the 5000 bp fragment is the original fragment. (b) The 3800 bp fragment can only be composed of the 2500 and 1300 bp fragments and the 3700 bp fragment composed of the 2500 and 1200 bp fragments. Accordingly, the restriction map can be established.

DNA sequencing. For other applications, for example, to compare restriction patterns in several individuals, cloning is too laborious and analysis is done directly on genomic DNA. Alternatively, interesting areas may be amplified by polymerase-chain-reaction (PCR) before amplification products are then analyzed by restriction. This reaction can then be analyzed again by normal gel electrophoresis, without any specific labeling.

Detection of Methylated Bases Since there are isoschizomers like *Hpa*II and *Msp*I (see above) that differ in their sensitivity towards a methylation within their recognition sequence, methylated bases can be detected by them.

As an example so-called CpG islands are found in several promoter regions of eukaryotic genes. They are sections of DNA where dinucleotide CpG is overrepresented. If a gene is transcriptionally inactive this is often connected to the methylation of cytosines in the CpG island of the gene. If in a CpG island not all restriction sites cleaved by *Msp*I are cleaved as well by *Hpa*II this is an indication of methylation and thus transcriptional inactivity of the respective gene. Restriction analysis has to be performed directly on genomic DNA and is detected by Southern blot analysis (Section 28.4.3).

The difficulty of DNA-methylation analysis is discussed in detail in Chapter 30.

Detection of Mutations and Restriction Fragment Length Polymorphisms (RFLP) Individuals within a population differ in the composition of their genomes. There exist highly conserved areas that are of high relevance for the carrier and which are nearly unchanged within the population or even among species (e.g., globin genes). A mutation of such a region may cause illness or the death of the carrier (e.g., sickle cell anemia as a mutation in globin genes). On the other hand, there are areas with several variants within a population, so-called polymorphisms.

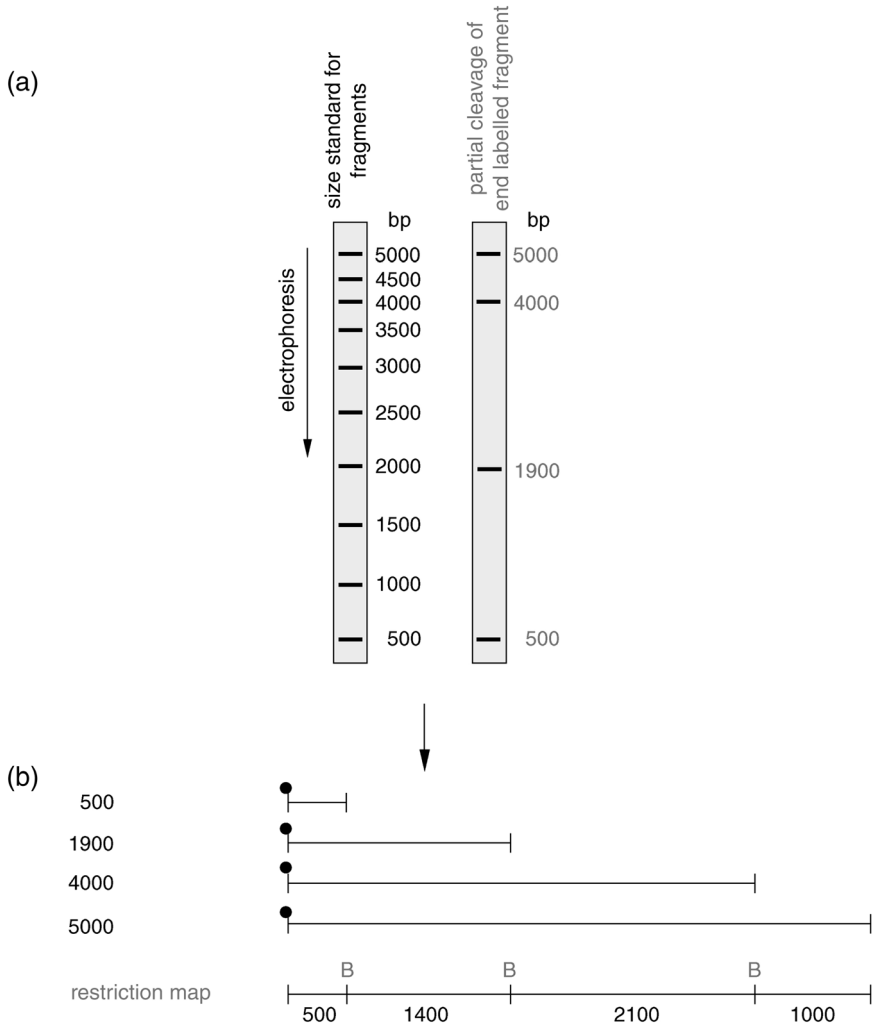


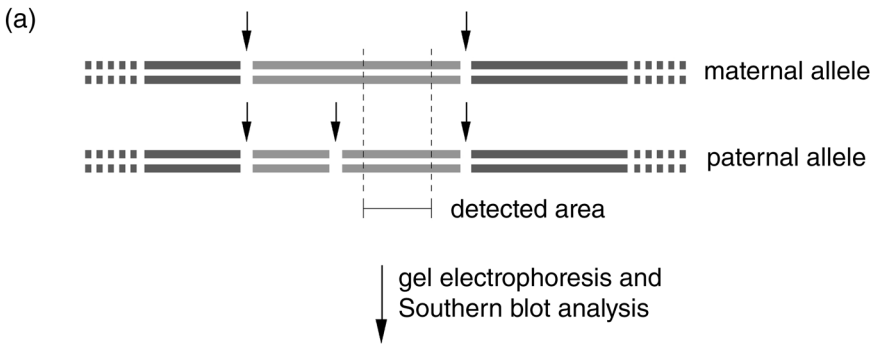
Figure 27.4 Partial restriction and end labeling. A 5 kb DNA molecule is labeled at one end, partially restricted by enzyme A, and the reaction products are separated by electrophoresis. Only the end-labeled fragments are detected (a). (b) The size of every fragment corresponds to the distance between a restriction site of enzyme A and the labeled end. The label is shown as a dot. The resulting restriction map is shown.

These differences in DNA sequence can be exchanges, deletions, or insertions of single bases or sections of DNA. These mutations can cause a change in length of a restriction fragment, or a restriction sequence can be deleted or inserted. If a polymorphic region can be detected by change of a restriction pattern this is called a RFLP. Hereby restriction analysis is either performed on genomic DNA in combination with Southern blot analysis (Section 27.4.3), and the interesting region is used as hybridization probe, or the region is amplified by PCR *in vitro* and restriction analysis is performed on the PCR product. Since every individual has two homologous copies of every DNA section, in the case of heterozygosity two kinds of restriction pattern will be detected when analyzing this RFLP, one representing the paternal and one the maternal allele (Figure 27.5). Figure 27.6 shows the heredity of a RFLP is over three generations.

Genetic Fingerprint A genetic fingerprint is based on the detection of highly variable RFLPs, which result in a restriction pattern that is highly characteristic for each individual. The basic causes for this are short, mostly two to three base pairs long, highly repetitive sequences, whose number of repetition is highly variable. This is helpful in identifying individuals, for example, as proof of paternity or in criminalistics (compare Section 27.2.1).

RFLPs as Genetic Markers, Section 36.1.2

Restriction Fragment Length Polymorphisms in Genetic Mapping In genetic mapping it is not the nucleotide sequence that is evaluated but the relative order of so-called genetic markers towards each other. This is done by gene linkage analysis. Possible genetic markers are blood groups and disease genes and also RFLPs. This is discussed in detail in Chapter 36.



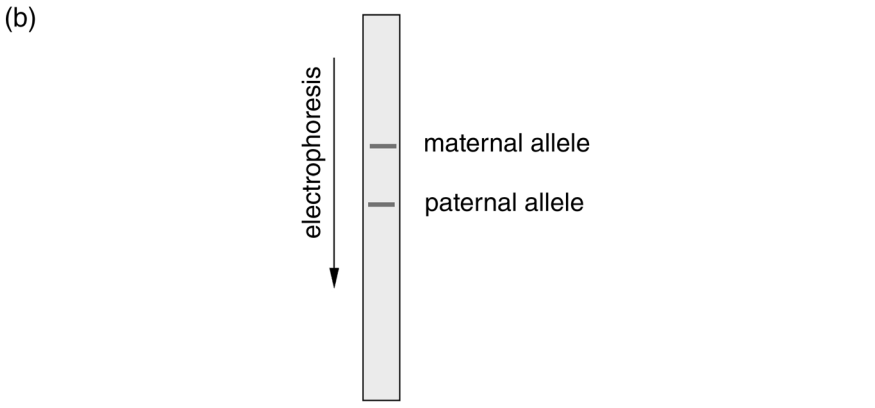


Figure 27.5 Detection of a RFLP by Southern blot analysis. (a) Homologous chromosomal sections of an individual containing a polymorphous region. Restriction sites are indicated by arrows. The area detected by the hybridization probe in Southern blot analysis (Section 27.4) is marked. (b) The respective result after restriction, gel electrophoresis, and Southern blot analysis. One restriction site is missing in the maternal allele, which results in detection of a longer restriction fragment. The shorter fragment corresponds to the paternal allele. Thus the respective restriction fragment is polymorphic in this individual.

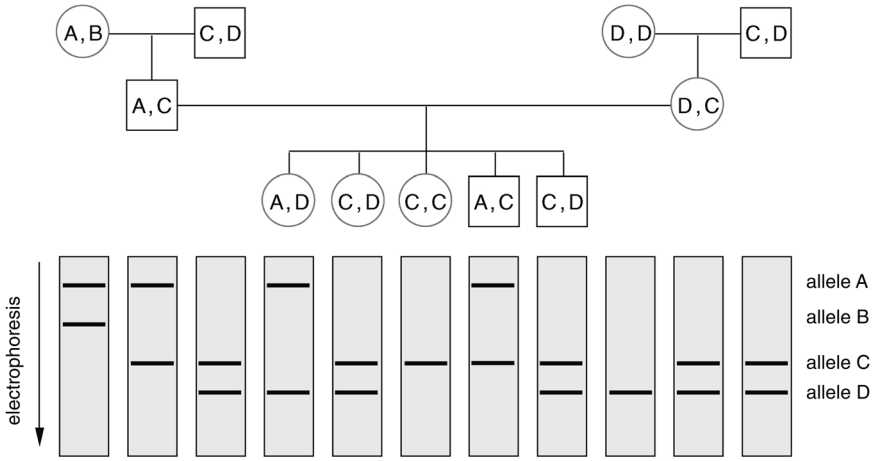


Figure 27.6 Heredity of a restriction fragment length polymorphism over three generations. In the family analyzed four alleles occur for the polymorphic region: allele A, B, C, and D. The heredity is in accordance with Mendel's laws. Most individuals are polymorphic for the restriction fragment analyzed, others have the same allele in both homologous areas.

27.2 Electrophoresis

Marion Jurk

Miltenyi Biotec GmbH Friedrich-Ebert-Straße 68 51429 Bergisch Gladbach Germany

Electrophoretic Techniques, Chapter 11

Electrophoresis is a most important method by which to analyze nucleic acids. Its advantages are obvious: electrophoresis can be performed in a very short time frame with low amounts of material. The necessary equipment and detection methods are in most cases very cheap and are easily available in every laboratory.

The underlying theoretical principles and the hands on work have similarities to, but also significant differences from, the electrophoretic separation of proteins. Like proteins, the separation of nucleic acids in an electric field is performed in a solid carrier material such as agarose or polyacrylamide. In contrast to proteins, nucleic acids are negatively charged within a very broad pH range. The negative charges are carried by the phosphate groups on the backbone of

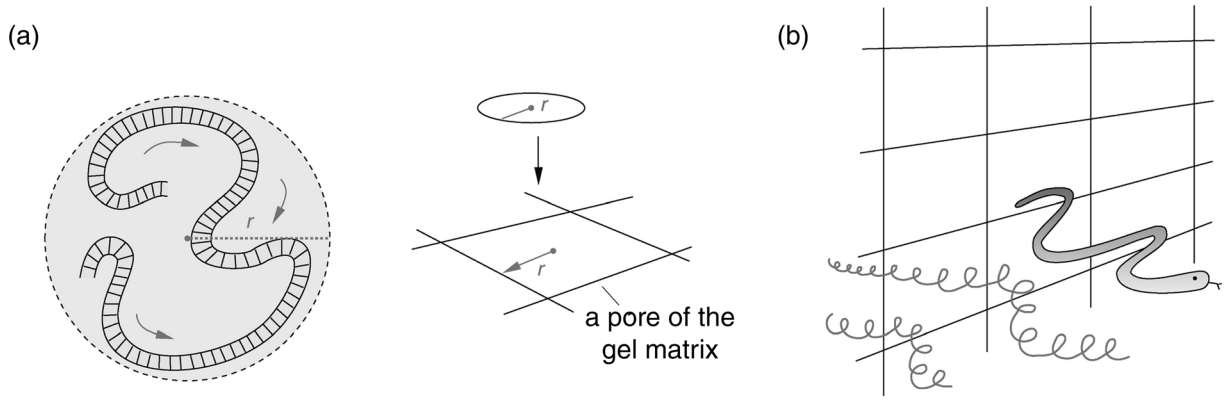


Figure 27.7 Theories explaining the movement of nucleic acids in the gel matrix. The Ogston theory (a) postulates a globular sphere for the nucleic acids. Its radius is defined by the length of the molecule and thermal agitation. The molecules migrate through the pores of the gel matrix if the diameter of the nucleic acid smaller than the average pore size. According to the reptation theory (b) the nucleic acids align themselves along the electric field and move snake-like through the gel matrix. Source: adapted according to Martin, R. (1996) *Gel Electrophoresis: Nucleic Acids*, Bio Scientific Publishers Limited, Oxford.

the nucleic acids. The migration of nucleic acids in the electric field towards the anode is therefore pH independent. Another notable difference from proteins is their constant charge density, meaning that the ratio of molecular weight to negative charge remains unchanged. There is no need to generate homogenous charge surfaces by SDS like it is the case with proteins.

The electrophoretic mobility, that is, the velocity of migration in the electric field (Chapter 12), is equal for all nucleic acids in free solution independent of their molecular weight. Differences in their mobility can only be measured in a solid gel matrix. The differences in migration velocity are caused solely by different molecule sizes.

The movement of nucleic acids in an electric field can be described by two theories (Figure 27.7). The migration of nucleic acids in reality could be seen as a "mixture" of these two theories.

The Ogston sieving effect is based on the assumption that nucleic acids in solution have a globular, spherical structure. The size of nucleic acids is described by the radius of the sphere that is theoretically occupied by the nucleic acid. The bigger the sphere the more often collisions with the gel matrix will occur. Migration of the nucleic acids in the field will then be slowed down. Very small fragments will not be slowed down by the pores of the gel matrix. Small fragments cannot be separated. According to the Ogston sieving theory, very big molecules with sphere sizes bigger than the pores of the gel should not be able to migrate at all. A second theory, the reptation theory aims to explain the migration of big nucleic acids in the electric field. The theory assumes that big nucleic acids can abandon their globular structure and align themselves in the electric field. The migration of the molecules occurs by moving one end of the molecule ahead through the matrix pores (end-to-end migration). The theory is called reptation owing to the snake-like movement of the nucleic acids. Size selection occurs because bigger molecules need more time to move than smaller ones. Both theories together can explain most of the phenomena observed in the electrophoresis of nucleic acids with sizes of 10 kb. The behavior of very large molecules cannot be explained by these theories and requires new model theories (Section 27.2.3).

27.2.1 Gel Electrophoresis of DNA

Agarose Gels The choice of carrier material depends mainly on the size and kind of nucleic acid to be analyzed. Agarose, a linear polysaccharide polymer, is the most important electrophoresis material for nucleic acids.

The migration velocity of DNA molecules is determined by several factors. The effective size of a nucleic acid is not only determined by its absolute mass but also depends on its form: superhelical (form I), open-circular (form II), double-stranded linear (form III), or single stranded.

Separation of Linear, Double-Stranded DNA Fragments Gel electrophoresis of linear DNA fragments (form III DNA) can be used to determine the size of the DNA reproducibly

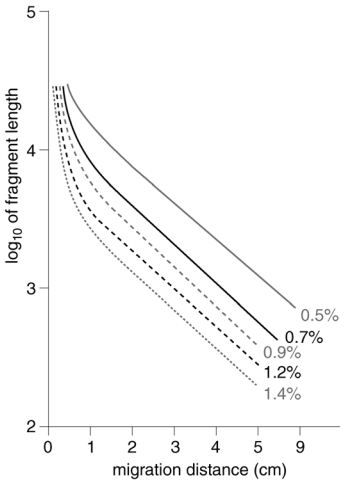


Figure 27.8 Relationship between the migration distance and fragment length at various agarose concentrations. The semi-logarithmic curves were created using length standards. The size of a fragment can be determined by its position. Buffer: 0.5 TBE/0.5 μg ml^{-1} ethidium bromide. electrophoresis at 1 V cm^{-1}; 16 h.Source: adapted according to Maniatis, T., Fritsch, F.F., and Sambrook, J. (1989) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor.

with good accuracy. There is a linear relation between the logarithm ($\log_{10}$) of the size (in bp) of the fragment and the migration distance (measured in cm in reference to the total distance) in an agarose gel (Figure 27.8). The migration velocity of linear DNA fragments is dependent on the concentration of the agarose, the applied voltage, composition of the running buffer, and the presence of intercalating dyes.

Linear DNA fragments can be separated by agarose gels spanning a broad range of fragment lengths (Table 27.4). Very small fragments (≤100 bp) migrate in 1–1.5% containing agarose gels at the same speed because the pores of the gels are bigger than the fragments. Separation of these small fragments becomes possible by increasing the agarose concentration. Small DNA fragments and oligonucleotides are usually separated using 2–3% agarose gels.

The migration speed of the fragments is proportional to the applied voltage. Big fragments migrate into the gel slowly if the voltage is too high – bigger fragments should therefore be separated at lower voltages. Good separation of fragments (≥2 kB) takes place when the applied voltage is less than 5 V cm^{-1}, with the distance between the electrodes being the influential parameter, not the length of the gel.

For the separation of DNA molecules a running buffer with Tris acetate (TAE) or Tris borate (TBE) is used. Fragments separated in TAE buffer can be better isolated from an agarose gel. The bands are usually sharper. A disadvantage of TAE buffer is its lower buffering capacity and lower stability during electrophoresis. If long electrophoresis times or high field forces are necessary, TBE buffer is used. Linear fragments migrate faster in TBE buffer (approx. 10% faster) than in TAE buffer. The separation capacity is similar in both buffer systems; however, superhelical DNA can be better separated in TBE buffer.

The ion concentration of the running buffer is of importance as well. Too low a concentration causes minimal electric conductivity and consequently the speed of migration of the nucleic acids is low. Too high a concentration results in a very high electric conductivity, which causes heating of the buffer. The DNA could possibly be denatured and the agarose melted. The presence of intercalating dyes influences the speed of the nucleic acids as well. The principle of intercalation is described in Section 27.3.1. Addition of ethidium bromide decreases the migration velocity of linear double-stranded DNA fragments about 15%.

Separation of Circular DNA The migration velocity of circular DNA form I (superhelical) or form II (open) depends mainly on the consistency of the agarose gel. Superhelical DNA migrates faster than linear DNA. Relaxed DNA molecules (form II) are slower than linear or superhelical DNA (Figure 27.9). The migration velocity of these three forms is influenced by the running conditions, concentration of the agarose, applied voltage, and choice of running buffer. The different forms can be identified by ethidium bromide staining.

Practical Considerations Agarose gels can be poured as vertical or horizontal gels. In most laboratories, the more practical, vertical gels are used. According to the size of the gel, there are mini, midi, and maxi gels. Mini gels have a very short distance for separation (6–8 cm) and are not suited for size determination of DNA fragments. They are used for a quick analysis of the quality of the DNA and for control of restriction digestion. Midi and maxi gels (approx. 20 or

Table 27.4 Coarse separation of DNA fragments at different agarose concentrations. Source: according to Sambrook, J. and Russell D.W. (eds) (2001) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Press, Cold Spring Harbor.

Agarose concentration (% w/v)	Optimal separation range of linear double-stranded DNA fragments (kb)
0.3	5–60
0.6	1–20
0.7	0.8–10
0.9	0.5–7
1.2	0.4–6
1.5	0.2–3
2.0	0.1–2

30–40 cm) are used for accurate DNA size determination and for the isolation of fragments. The separation distance and loading capacity of the gel are much higher. The DNA is loaded onto the gel using a so-called loading buffer. They increase the density of the DNA solution (using Ficoll, glycerol, or saccharose). The DNA solution sinks into the gel pockets and does not diffuse to the running buffer. The loading buffer usually contains negatively charged dyes indicating the migration during electrophoresis. The most commonly used dyes are bromophenol blue and xylene cyanol. Bromophenol blue migrates in an agarose gel, depending on the exact conditions, like a linear DNA fragment of approx. 300 bp.

An important means for the determination of DNA fragment sizes are DNA standards. DNA standards are commercially available and contain DNA fragments of defined sizes. They are separated together with the DNA fragments of interest. Using the known sizes of the DNA standard, the unknown size of the DNA fragment can be determined. For an accurate size determination it is very important that the DNA standard and unknown fragment are loaded as similar amounts and with similar buffer conditions. DNA standards can be produced by restriction cleavage of plasmid or phage DNA. A common DNA standard is the 1 kb DNA ladder (Figure 27.10). There are also standards available for small size DNA fragments, for example, the 100 bp ladder consisting of DNA fragments that differ exactly by 100 bp. The choice of standard depends on the size of the expected DNA fragments.

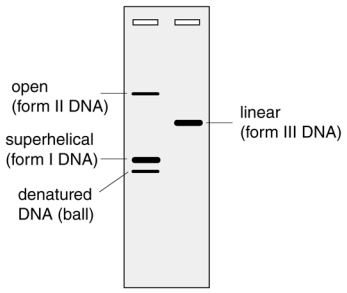
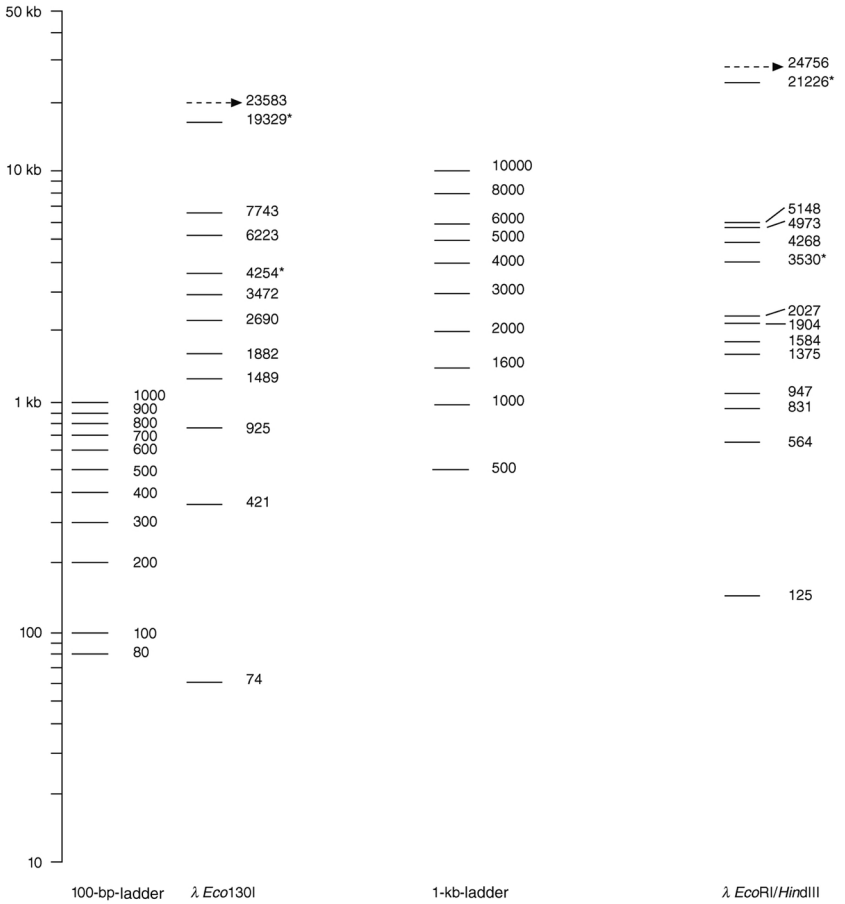


Figure 27.9 Migration behavior of superhelical, open, linear, and denatured DNA in an agarose gel. The migration properties of superhelical DNA can be influenced by the ethidium bromide concentration.

Denaturing Agarose Gels Single-stranded DNA forms intramolecular secondary structures and intermolecular aggregates very easily. These structures influence the migration behavior in a gel. For the size determination of single-stranded DNA, denaturing agarose gels are employed where the electric mobility of the DNA depends solely on the molecular size.



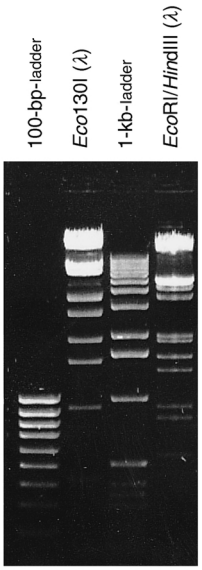


Figure 27.10 Various commonly used DNA length standards. The λ-DNA marker can be generated by digestion with certain restriction enzymes (Section 27.1). Source: Photography by Dr. Marion Jurk.

Alkaline agarose gels are used to determine the synthesis efficiency of first and second strand synthesis of cDNA and to test the nicking activity of enzyme preparations. Sodium hydroxide is used as denaturing agent. The agarose needs to be dissolved in water because the addition of hot sodium hydroxide hydrolyzes its polysaccharide structures. Since ethidium bromide does not bind to the DNA at high pH, the electrophoresis is performed in the absence of ethidium bromide.

Low Melting Agarose and Sieving Agarose Derivatization of agarose by the introduction of hydroxyethyl groups into the polysaccharide chains results in an agarose moiety with different properties. This low melting agarose is also heated and will gel when cooling down. However, its melting point is reduced. This property is used when isolating DNA fragments out of agarose gels (Section 27.5.3). The migration velocity of DNA in low melting agarose gels is higher, the separation and loading capacity is lower.

The properties of sieving agarose are similar to those of low-melting agarose. Sieving agarose is used especially for the separation of small DNA fragments. Both types of agarose should not be used at concentrations below 2% for reasons of stability. Recommended concentrations are 2–4%.

Gel Media for Electrophoresis, Section 11.3.2

Polyacrylamide Gels The properties of polyacrylamide and the definitions of concentration and degree of crosslinking have been introduced in Section 12.3.2. Electrophoresis of DNA in polyacrylamide gels (abbreviated as PAGE, polyacrylamide gel electrophoresis) can be performed under native or denaturing conditions depending on the scope of application. Polyacrylamide gels (these gels as well as agarose gels are colloquially called slab gels) are poured as horizontal gels between two glass plates. Advantages and disadvantages of polyacrylamide and agarose gels are listed in Table 27.5.

Non-denaturing Gels for Analysis of Protein–DNA Complexes Native, non-denaturing gels are used for electrophoretic mobility shift assays (EMSAs). With this method, protein–DNA complexes can be separated from free DNA. Large DNA–protein complexes are retarded in the gel by the cage effect. This method is described in detail in Chapter 32.

Non-denaturing PAGE of Double-Stranded DNA Native polyacrylamide gels yield a higher resolution than agarose gels (Table 27.6) together with a higher loading capacity. This is used for the purification and isolation of double-stranded DNA fragments (<1000 bp). The gels are prepared in TBE buffer and stained with ethidium bromide after the electrophoresis. Isolation of DNA from the gels is described in Section 27.4.

Abnormal Migrational Behavior of DNA Fragments in Native Polyacrylamide Gels (Curvature) Since polyacrylamide gels have a higher resolution, any abnormal migration properties of certain DNA fragments can be analyzed by PAGE. Small DNA fragments (350–700 bp) with certain sequences display a reduced migration compared to their actual size

Table 27.5 Comparison of agarose and polyacrylamide gels.

Agarose gels	Polyacrylamide gels
Advantages	
Less technical effort, easy handling, faster	Higher resolution for DNA within a certain length range (>1000 bp)
Greater resolution range	Loading capacity higher without loss of resolution
DNA can easily be isolated	Purification yields DNA of high quality
DNA can easily be stained	
Capillary and vacuum blotting	
Disadvantages	
Bands are more diffuse and broader	Difficult to pour, greater technical effort
Resolution of smaller fragments is lower	No capillary or vacuum blotting possible
Isolated DNA fragments can contain impurities	Lower separation range

Table 27.6 Separation range of native polyacrylamide gels. The ratio of acrylamide to N,N′-methylene-bisacrylamide is 29 : 1.

Acrylamide concentration (%)	Separation range (bp)	Migration of bromophenol blue in native gels (bp)
3.5	100–2000	100
5.0	100–500	65
8.0	60–400	45
12.0	50–200	20
15.0	25–150	15
20.0	5–100	12

(number of base pairs) under certain electrophoresis conditions. This effect is thought to be due to conformational changes in the DNA, such as kinks or bends. The abnormal migration behavior is more pronounced at higher polyacrylamide concentrations, higher Mg^{2+} concentrations, or at lower temperatures. An increase in temperature or concentration of Na^{+} ions results in opposite effects.

Native PAGE of Single-Stranded DNA (SSCP) These gels are usually used to analyze changes within genomic DNA for certain disease indications. For the determination of various genetic mutations, methods are needed that allow many patient samples to be run at the same time. Sequencing of the individual genomic DNA would be far too expensive and time consuming. The commonly used SSCP (single-stranded conformational polymorphism) method is based on the observation that single-stranded DNA molecules with different sequences assume different conformations. The double-stranded DNA fragments to be analyzed are denatured using formaldehyde and are loaded onto a native polyacrylamide gel. The isolated single-stranded DNA strands will assume individual conformations resulting in different migrational behavior (Figure 27.11). Gene sections of individuals

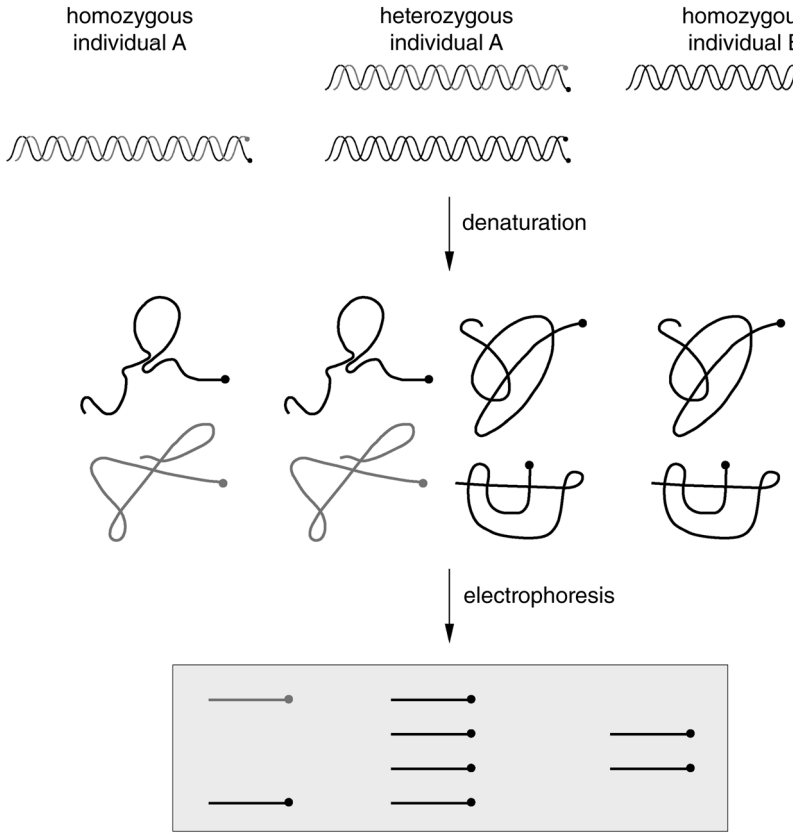


Figure 27.11 Schematic principle of SSCP analysis. The DNA fragments are denatured in the presence of formamide by heat. The resulting single-stranded molecules assume a certain conformation according to their sequence and base composition. The single-stranded molecules are separated using native polyacrylamide gel electrophoresis. The migration properties differ according to the different conformations. From the characteristic gel pattern, homozygous and heterozygous individuals with genes containing certain point mutations in certain genes can be identified. Source: adapted according to Martin, R. (1996) *Gel Electrophoresis: Nucleic Acids*, Bio Scientific Publishers Limited, Oxford..

Table 27.7 Separation of oligonucleotides in denaturing polyacrylamide gels. The ratio of acrylamide to N,N′-methylene-bisacrylamide 19 : 1.

Acrylamide concentration (%)	Separation range (in nt; nt = nucleotide)
20–30	2–8
15–20	8–25
13.5–15	25–35
10–13.5	35–45
8–10	45–70
6–8	70–300

containing point mutations will have a different conformation and consequently result in a different migration speed. It is essential to perform a native PAGE as denaturing gels would result in a uniform separation according to the size of the fragments but not according to sequence.

Using SSCP various different DNA samples can be analyzed for point mutations simultaneously. The first SSCP analysis was performed using restriction generated fragments of DNA following Southern blot analysis (Section 27.4.3). In more recent approaches, the gene sections of interest are amplified using PCR and radioactive labeled with no follow up detection necessary. The analysis is performed under the assumption that mutations will behave differently than the original fragment. A negative result is not irrevocable proof that there are no point mutations present in the analyzed gene section.

Denaturing PAGE of single-stranded DNA or RNA is used in various fields of applications due to the very exact separation of the single-stranded molecules (Table 27.7). The most common denaturing agent is urea, but formaldehyde is also used. Alkaline reagents cannot be used with polyacrylamide gels because the gel matrix will be destroyed. The gels are usually polymerized in the presence of 7 M urea, the running buffer is TBE. The loading buffer is usually formaldehyde to denature the probes (Section 27.2.2). Single-stranded DNA or RNA migrates in this type of gel independent of its sequence and, therefore, the separation of DNA molecules differing in size by only in one nucleotide is possible. These gels are therefore used for sequencing, S1 nuclease analysis, and RNAse protection experiments. The gels are also used for the DNA fingerprinting method.

DNA Fingerprinting DNA fingerprinting is applied for the lineage analysis of genomic DNA. The technique is used in forensic analysis and for zoological studies, and also for paternity testing. DNA minisatellites, repetitive variant repeats in the genomic DNA of 10–100 bp, are inherited to a similar extent from both parents. The distribution and cleavage behavior are unique for each individual. The genomic DNA is cut by restriction enzymes and separated on denaturing polyacrylamide gels. The gels are blotted to a membrane (Southern blotting, Section 27.4.3) and hybridized using specific probes that recognize the minisatellite DNA. Fingerprinting can also be performed using PCR with random primers. Short DNA fragments are synthesized by PCR in the presence of radioactive labeled nucleotides. Two individuals will differ in their spectra of synthesized DNA fragments when run on a denaturing polyacrylamide gel with high resolution. The electrophoretic methods have been replaced but will eventually be completely replaced by the next generation of sequencing methods.

Oligonucleotide Purification Denaturing polyacrylamide gels are often used for the purification of synthetic oligonucleotides or single-stranded DNA. Oligonucleotides with n bases can be separated from oligonucleotides with $n - 1$ bases, yielding a population of nucleic acids with uniform length. The oligonucleotides are detected by fluorescent quenching. The gel is laid on a thin-layer chromatography (TLC) plate and irradiated with UV light (long wavelength). The TLC plate fluoresces upon excitation, except for the parts where the oligonucleotides are located. The oligonucleotide can be identified as dark bands and excised from the gel. If the oligonucleotides need to be isolated from the gel, the excitation time should be kept to a minimum to avoid damage to the nucleic acids or crosslinking to the gel matrix.

When this method is chosen for oligonucleotide purification the oligonucleotides should not contain modifications that interact with the polyacrylamide matrix.

27.2.2 Gel Electrophoresis of RNA

Similar to single-stranded DNA, single-stranded RNA forms secondary structures by intra- and intermolecular base pairing. These different conformations behave differently during gel electrophoresis. An exact and reproducible analysis of RNA is only possible using denaturing gel electrophoresis. In denaturing gels the hydrogen bridges are destroyed and all RNA molecules will be separated according to their molecular weight.

The electrophoresis of complex RNA mixtures (e.g., for Northern blotting, Section 27.4.4) is performed using denaturing 1–1.5% agarose gels. Smaller RNA fragments are separated like oligonucleotides using denaturing PAGE. For a rapid analysis of the RNA (e.g., for reasons of quality control), non-denaturing TBE gels can be used.

The denaturing reagents employed are usually dimethyl sulfoxide/glyoxal or formaldehyde for agarose gels, and urea for polyacrylamide gels.

Since RNA is subject to nuclease digestions and hydrolysis by acids or bases, the experimental set-up for RNA electrophoresis has to be modified compared to DNA electrophoresis. The same precautionary measures as for isolation of RNA have to be taken for the electrophoresis of RNA. For example, the electrophoresis chamber needs to be cleaned carefully and only RNAse-free water should be used.

Formaldehyde Gels The denaturing effect of formaldehyde is based on the formation of so-called Schiff bases between the aldehyde functional group and the amino group of the adenine, cytosine, and guanine bases. Consequently, the amino groups of the nucleobases cannot form hydrogen bonds for the formation of secondary structure or aggregates. The agarose gel usually contains 1.1 % formaldehyde. For longer separations (overnight), the formaldehyde content must be increased. As formaldehyde is toxic, electrophoresis should be performed under a fume hood.

Schiff bases are generated by the reaction of primary amino groups with aldehydes and water is released. An imine bond is formed.

Since formaldehyde also interacts with the amino groups of Tris (Tris(hydroxymethyl) aminomethane), a different running buffer has to be used. This is usually a mixture containing 3-*N*-morpholino-1-propane sulfonic acid (MOPS) and sodium acetate. The RNA has to be denatured before loading onto the gel in the presence of formaldehyde using formamide and MOPS. The formamide destroys the base pairing of the RNA, allowing the formation of Schiff bases between the formaldehyde and the amino groups.

Formamide can be contaminated by ions like ammonium formate, which can hydrolyze the RNA. Formamide is therefore deionized using ion exchange chromatography. Since MOPS possesses a very high buffering capacity, there is no need to replace or recycle the running buffer during the run as it is the case with glyoxal gels. If the gel needs to be blotted afterwards, the formaldehyde needs to be removed before blotting. Otherwise the amino groups of the bases will not be available for hybridization with the probes.

Glyoxal gels yield sharper RNA bands than formaldehyde gels, which is an advantage for blotting. Glyoxal binds to the guanine residues at neutral pH and prevents base-pairing of the RNA. In contrast to formaldehyde gels, glyoxal is only added before loading and is not added to running or loading buffer. The RNA is denatured in the presence of 1 M glyoxal and sodium phosphate and dimethyl sulfoxide (DMSO) at 50 °C. Sodium phosphate acts as buffer and DMSO destroys the inter- and intramolecular hydrogen bonds, enabling the glyoxal to react with the guanine residues. Glyoxal is easily oxidized to glyoxylic acid, which hydrolyzes RNA. Contaminating glyoxylic acid has to be removed by ion exchange before use of glyoxal in gel electrophoresis. Glyoxal reacts with ethidium bromide; consequently, the separation is performed in the absence of the intercalating reagent. Above pH 8.0, the glyoxal dissociates from the RNA. To avoid pH gradients during electrophoresis, the running buffer has to be replaced or recycled using a pump.

RNA Standards Cytoplasmic RNA of eukaryotic cells consists of approx. 95% of ribosomal RNA (rRNA). Ribosomal RNA consists of 28S, 18S, and 5S rRNA. RNA preparations of high quality display two sharp, clearly separated bands in an agarose gel (Figure 27.12) that can be used as internal standards. The exact size of the ribosomal RNA depends on its origin: for human rRNA the lengths were determined to be 5.1 kb for 28S and 1.9 kb for 18S rRNA. Other length standards are commercially available or can be generated by *in vitro* transcription of DNA fragments of defined length.

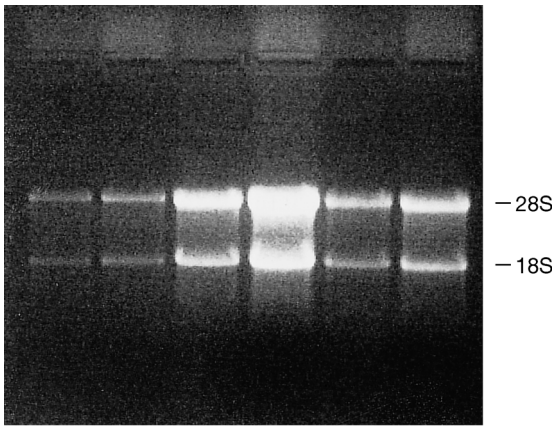


Figure 27.12 Migration of cytoplasmic RNA. High quality RNA preparations should generate clearly visible bands of the 28S and 18S of ribosomal RNA and should barely be degraded. These bands can be used as internal length standards. Source: Photography by Dr Marion Jurk.

27.2.3 Pulsed-Field Gel Electrophoresis (PFGE)

Principle High molecular weight nucleic acids cannot be separated by regular gel electrophoresis. They all possess the same so-called limiting mobility. This effect cannot be explained by the reptation theory; several other theories have been put forward to explain the observed phenomenon. One model postulates that the DNA molecules act as rigorous entities, whereby no separation effect can be obtained. Another model describes the movement of high molecular weight DNA as similar to movements in solution, where no separation effect can occur. The formation of loop structures could also explain the limiting mobility effect.

Pulsed-field electrophoresis (PFGE) uses, instead of a continuous electric field, a pulsed field with changing directions of the electric fields. DNA molecules assume a relaxed globular shape in free solution (without an electric field). When an electric field is applied, the molecules align themselves according to the electric field and will move towards the anode (according to reptation theory). On removing the electric field, the molecules will again resume the relaxed globular state. By applying an electric field with a different direction, molecules will realign according to the new field. If the direction of the electric field is changed again, the molecules have to realign again. Relaxation and alignment of the molecules according to the electric field depends on the size of the molecule. Larger molecules need more time for relaxation and alignment than smaller molecules. The time needed for movement along the electric field is shorter for larger molecules than for smaller ones. The sum of all applied field vectors yields the direction of movement of the DNA molecules. The separation principle of large DNA molecules here is based on the time the molecule needs to align according to the applied electric fields.

Within PFGE, several techniques are condensed, most of them differing in direction and sequence of the electric pulses.

Field inversion gel electrophoresis (FIGE) uses two electric fields with opposing directions. Migration is achieved because the duration and amplitude of the forward pulse is larger (Figure 27.13). The method can separate molecules within a broad size range with high resolution.

The CHEF (contour-clamped homogenous electric field) method is a more commonly used PFGE method. The electrodes are arranged in a hexagon around the agarose gel (Figure 27.14).

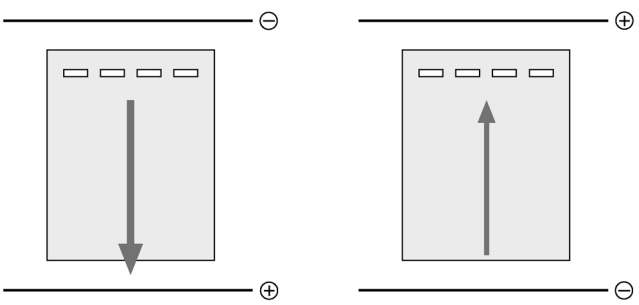


Figure 27.13 Principle of field inversion gel electrophoresis (FIGE). Two alternating electric fields with directions differing by 180°. The migration direction to the anode is determined by a longer or stronger pulse in this direction.

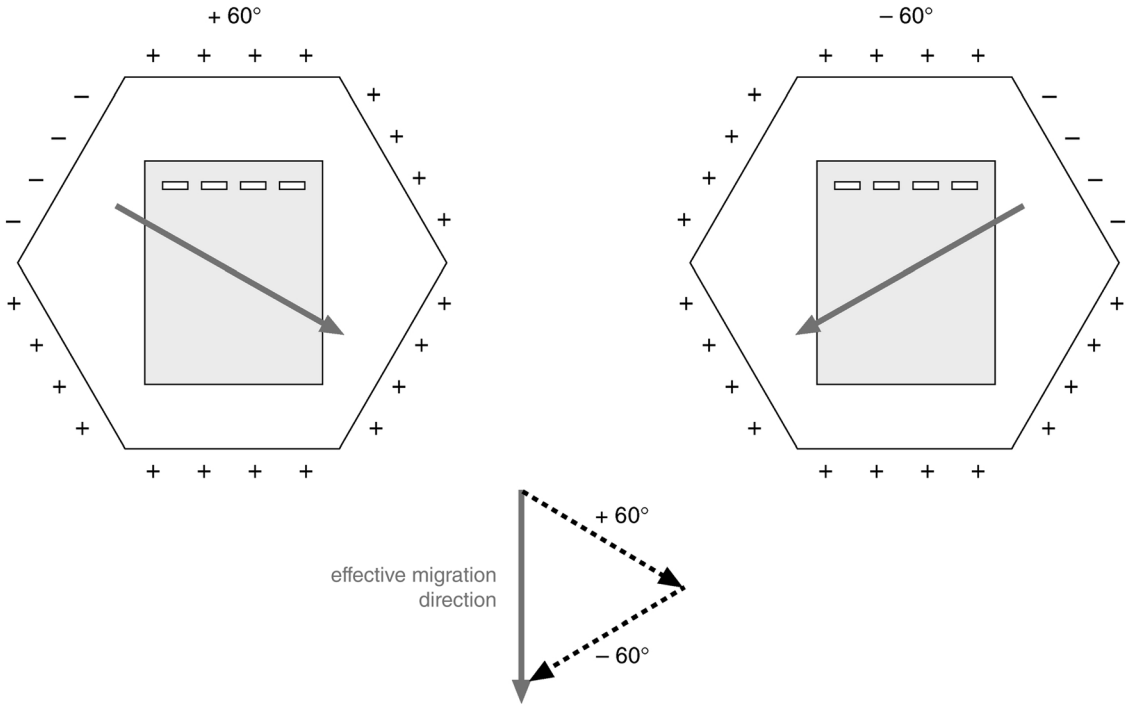


Figure 27.14 Principle of the contour-clamped homogenous electric field (CHEF) method. The pulses are applied in different directions. The migration of the DNA is determined by the sum of all applied field vectors and is, as displayed, a zigzag pattern.

The electric field is applied in such a way that the field vectors are aligned at angles of −60° and +60° relative to the vertical axis of the gel. The resulting movement of the nucleic acids resembles a zigzag pattern. The angle, field strength, and duration of pulses can be varied. With this method molecules up to 2000 kb can be separated.

An improved version of the CHEF method is PACE (programmable autonomously controlled electrode). Twenty-four hexagonally arranged electrodes can perform any desired pulse sequence. Improvements of pulse sequences result in optimal resolution and separation properties of the gel.

Applications For the separation of high molecular weight DNA the integrity of the nucleic acids is of upmost importance. To avoid any destruction of the DNA, the material is packed into agarose blocks before lysing the cells with detergents and proteinase K. High molecular weight DNA is isolated within the agarose block by incubating the agarose in the respective buffers. The agarose blocks are then applied to the gel pockets of the PFGE gel (usually 1% agarose gels).

With PFGE, higher voltages are applied, resulting in a temperature increase of the running buffer due to the higher currents. The running buffer is diluted TBE (0.5 × TBE) with the addition of glycine. Glycine increases the mobility of DNA without influencing the current. To avoid pH and temperature differences, the buffer is recycled during electrophoresis. The electrophoresis is usually performed in the absence of ethidium bromide, but when separating molecules with sizes smaller than 100 kB the addition of ethidium bromide can increase the resolution efficiency because the dye influences the reorientation of the DNA molecules.

The length and type of pulse sequences is very variable for the different types of PFGE and need to be optimized according to the individual conditions. Pulses are between 5 and 1000 s, field strength is usually between 2 and 10 V cm^{-1}. Running times can vary from 10 h up to several days.

Length standards can be high molecular nucleic acids like the genomic DNA of bacteriophage T7 (40 kb), T2 (166 kb), or G (756 kb). Ligation of bacteriophage lambda DNA yields an optimal length standard (Figure 27.15) with multiples of the lambda-DNA (48.5 kb)

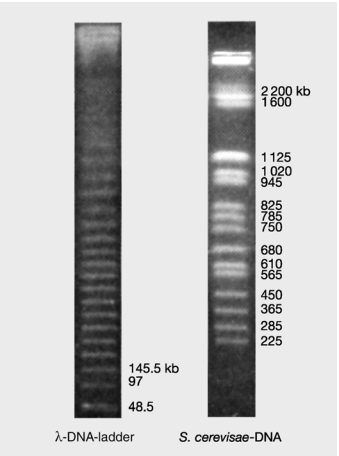


Figure 27.15 Common length standards for PFGE gels. The λ-DNA ladder can be generated by ligation of λ-phage DNA. Source: with kind permission of Bio-Rad, Munich.

PFGE is commonly used for the analysis of pathogens from food or clinical isolates. Different strains of certain bacteria (e.g., *Listeria monocytogenes*) can be analyzed according to its origin. Although with PFGE a resolution of 5 Mb is possible, PFGE cannot resolve human chromosomes (>50 Mb). However, using restriction enzymes, mapping and analysis of the human genome is possible. Rare cutters (e.g., NotI, NruI, MluI, SfiI, XhoI, SalI, SmaI (see Section 27.1)) are used. The PFGE gels are then blotted and hybridized. Physical maps of the human genome by PFGE are used for genomic fingerprinting and to analyze chromosomal deletions or translocations. The whole genome mapping method is also used for subtyping pathogenic strains.

27.2.4 Two-Dimensional Gel Electrophoresis

Two-dimensional gel electrophoresis (2D gel electrophoresis) is necessary when information obtained by one-dimensional electrophoresis is not sufficient or clear-cut. The high resolution of 2D gels is achieved by repeating the electrophoresis under completely different conditions. Complex nucleic acid mixtures can be separated that cannot be achieved by a single electrophoresis.

The nucleic acid mixture is first separated through a standard electrophoresis where the nucleic acids are separated by their molecular weight (first dimension). The gel lane containing the separated nucleic acids is cut out of the gel and applied to a second gel where the nucleic acids are separated under different conditions (second dimension, Figure 27.16). Usually, the electric field of the second dimension is perpendicular to that of the first dimension. Two-dimensional electrophoresis can be applied in the analysis of RNA and DNA.

Two-Dimensional Electrophoresis of RNA The electrophoresis conditions for RNA differ regarding the concentration of urea, polyacrylamide, and the pH of the two dimensions (Table 27.8). By performing an electrophoresis in the presence (first dimension) or absence (second dimension) of urea (urea shift), the nucleic acid is first separated according to its size (first dimension) and then according to its conformation (second dimension). A change in concentration of polyacrylamide within the two dimensions can separate the RNA molecules by interaction with the gel matrix. Molecules with different conformations can display a similar migration behavior at a given polyacrylamide concentration but will differ when the pore size of the gel is changed. In addition, the concentration of urea, pH, and pore size can be changed simultaneously. The net charge of the RNA molecules will be influenced by lower pH, so that not all RNA molecules are negatively charged. As certain bases are protonated more easily at lower pH the net charge of the whole RNA molecule depends on the sequence. The second dimension is then performed under conditions that separate the nucleic acids according to their molecular weight.

Two-Dimensional Electrophoresis of DNA The 2D electrophoresis of DNA can be used for genome mapping, whereby the DNA is first cut with one restriction enzyme and separated. The separated fragments are then cut with a second enzyme and a second electrophoresis is performed. Fragments that are not cut by the second enzyme will be found on the diagonal of the gel; only

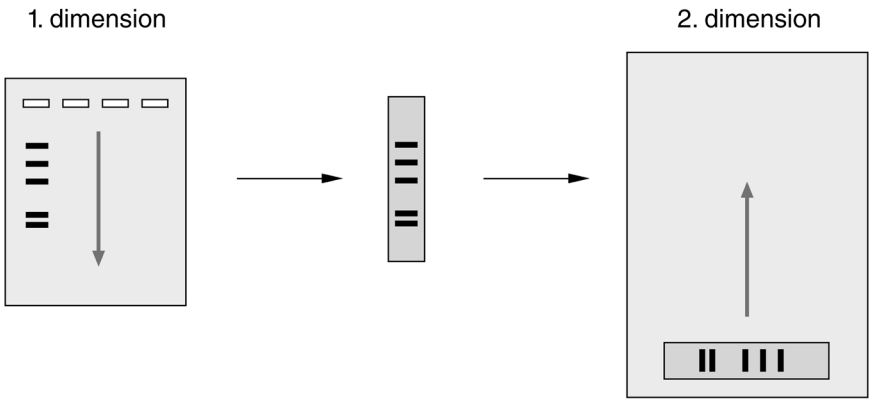


Figure 27.16 Principle and practical application of two-dimensional gel electrophoresis. The gel lane is isolated after first dimension electrophoresis and loaded to a second dimension gel. The direction of the second electrophoresis is rectangular with respect to the first dimension.

Table 27.8 Experimental conditions for the 2D gel electrophoresis of RNA molecules.

	First dimension			Second dimension		
	%PAA[a)]	pH[b)]	Urea (M)	%PAA[a)]	pH[b)]	Urea (M)
Urea-shift	X	Neutral	5–8	X	Neutral	0
	X	Neutral	0	X	Neutral	5–8
Concentration shift	X	Neutral	0	2X	Neutral	0
	X	Neutral	4	2X	Neutral	4
pH/urea/concentration	X	Acidic	6	2X	Neutral	0

a) X represents a certain polyacrylamide concentration.
b) The pH range of neutral electrophoresis is 4.5–8.5. Acidic electrophoresis occurs at a pH below 4.5. Neutral gels are usually run at pH 8.3, acidic gels at pH 3.3. A typical polyacrylamide concentration is 10–15%.

fragments cut by the second enzyme will run differently. With new methods, like next generation sequencing, these methods are becoming less important.

Two-dimensional electrophoresis has also been applied in mapping the origins of replication and in the analysis of topoisomers of superhelical DNA. The curvature of DNA can also be analyzed by 2D gel electrophoresis.

Topoisomers forms of DNA molecules of identical length and sequence that differ only in their linking number.

Temperature Gradient Gels A distant variant of the 2D gels is temperature gradient gel electrophoresis (TGGE), where the second dimension is temperature. The electrophoresis is performed in one direction and a temperature gradient is applied perpendicular to the gel (Figure 27.17). For this method one sample has to be loaded to the whole width of the gel. With increasing temperature the DNA is converted into the denatured state. Melting of DNA is a cooperative process accompanied by a drastic reduction of electric mobility. The process is strongly dependent on the sequence of the DNA fragments, since A/T rich regions melt at lower temperatures. Temperature gradients are applied for mutational analysis as this method can resolve single base changes. Using parallel TGGE (temperature gradient parallel to the electric field), many different samples can be analyzed simultaneously.

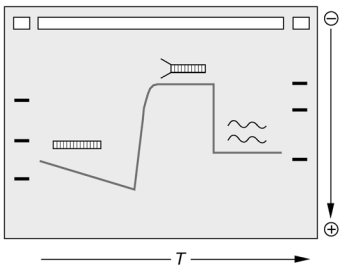


Figure 27.17 Principle of temperature gradient gels. The DNA is loaded onto the whole gel width and a temperature gradient is applied perpendicularly to the electric field.

For denaturing gradient gel electrophoresis (DGGE), the temperature gradient is replaced by a chemical-based gradient with increasing concentrations of denaturing reagents in the opposite direction to the electric field. Double-stranded DNA fragments are separated according to their melting properties. Both methods, TGGE and DGGE, are applied in heteroduplex analysis and for the analysis of microorganisms in environmental analytics.

27.2.5 Capillary Gel Electrophoresis

Capillary gel electrophoresis (CGE) is mainly applied for analysis of nucleic acid. The advantages of this method lie in its rapidity, small sample volume, higher sensitivity, and high resolution. The theory of CGE is covered in Chapter 12. The separation principle is the migration of negatively charged nucleic acids in an electric field. The separation is performed in a capillary (50–100 µm in diameter, approx. 20–50 cm long) utilizing the sieving effect of the gel matrix. An important difference to the already described slab gels is that in CGE non-crosslinked gels can also be used as sieving material. Crosslinked gels (also referred to as chemical gels) usually consist of polyacrylamide and can be used in the capillary for 30–100 runs. Non-crosslinked gels (referred to as physical gels) can be easily replaced after each run from the capillary by applying pressure, allowing each run to be performed under reproducible conditions with fresh material. Polymers used for physical gels are hydroxypropylmethyl cellulose (HPMC), hydroxyethyl cellulose (HEC), poly(ethylene oxide) (PEO), polyvinylpyrrolidone (PVP), or linear polyacrylamide. The running buffer is similar to that used with slab gels, TBE. For denaturing conditions urea is added. The samples (1–2 µl) are loaded by electrokinetic injection or with pressure (Chapter 12). Injection of the sample and migration in the capillary are strongly dependent on the salt concentration – in most cases the probes are desalted before loading. During electrophoresis high voltages are applied (1–30 kV). The nucleic acids can be detected by UV light in the presence of fluorescent dyes (OliGreen®,

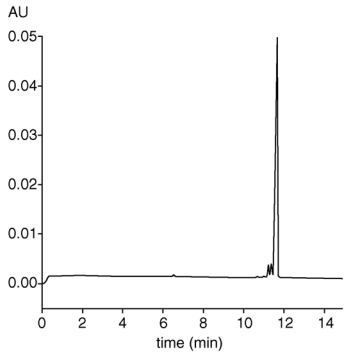


Figure 27.18 CGE electropherogram of an oligonucleotide. An oligodeoxyribonucleotide 23 bases long was analyzed by denaturing capillary electrophoresis. The main peak is the full-length product and the lower peaks are contaminations with failure sequences ($n-1$, $n-2$). By determination of the area under the curves, the ratio of the different products can be calculated. Source: with kind permission of Dr. Bernhard Noll, Qiagen GmbH.

SYBR®Green, Section 27.3). The UV-opaque polyimide layer stabilizing the capillary has to be removed at the detection site (Chapter 12).

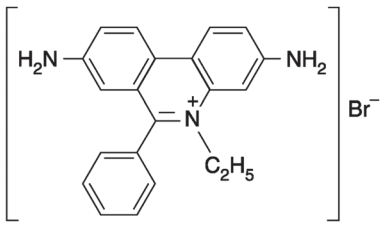
An important application of CGE is the quality control of synthetic oligodeoxy- and oligoribonucleotides as CGE allows resolution of one nucleotide difference. Oligonucleotides of up to 100 nucleotides can be separated and failure sequences ($n-1$, $n-2$, etc.) can be detected. Using the area under the curve method, the ratio of the separated nucleic acids and hence the purity of the material can be determined. Figure 27.18 shows an electropherogram of an oligonucleotide.

CGE is also used in nucleic acid sequencing. The sequence reactions labeled with four different dyes are separated by CGE and detected with laser-induced fluorescence (LIF). Conformational polymorphisms (SSCP and HA) can also be analyzed by CGE.

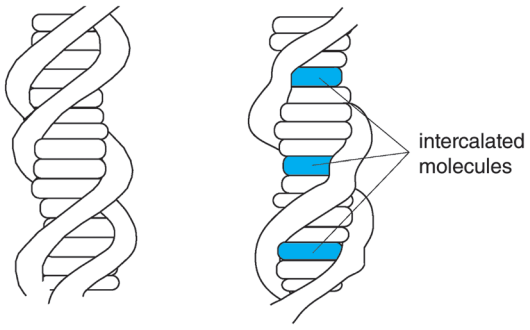
27.3 Staining Methods

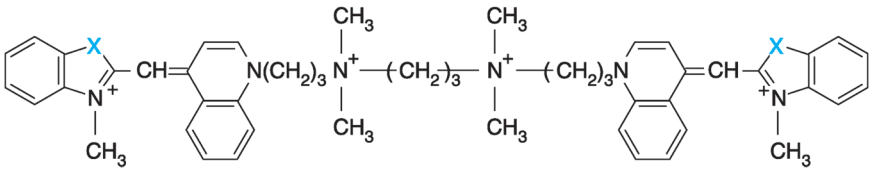
27.3.1 Fluorescent Dyes

Ethidium bromide Ethidium bromide (3,8-diamino-5-ethyl-6-phenylphenanthridinium bromide) is an organic dye with a planar structure that intercalates to DNA (Figure 27.19). The aromatic rings can interact with the heteroaromatic rings of the nucleobases. Single-stranded DNA or RNA also intercalates ethidium bromide but to a much lesser extent. The intercalated dye is excited by UV light (254–366 nm) and emits orange–red light (590 nm). Binding of the dye to DNA increases the fluorescence (increased quantum yield) so that the staining of the DNA is also



ethidium bromide





X = O: *YOYO-1*
X = S: *TOTO-1*

Figure 27.19 Chemical structure of ethidium bromide. Ethidium bromide intercalates preferentially into double-stranded DNA and interacts with the planar heterocyclic rings of the nucleobases. Newly developed fluorescent dyes that irreversibly stain the DNA are TOTO-1 and YOYO-1.

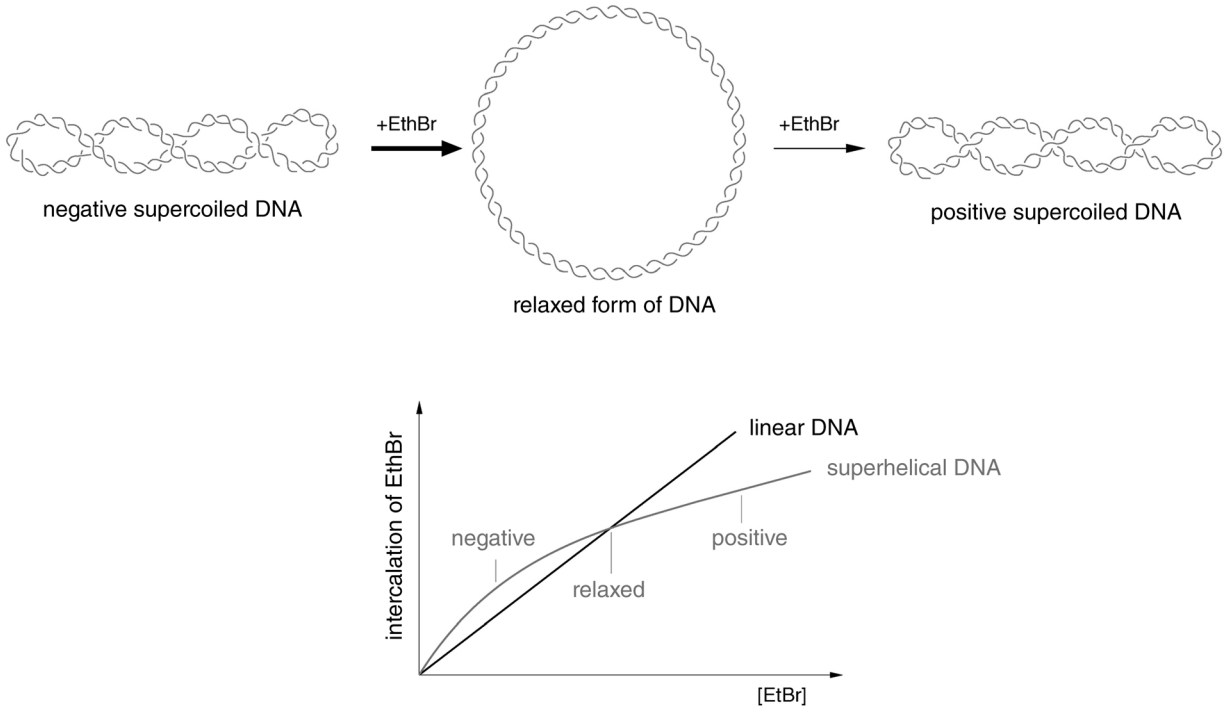


Figure 27.20 Changes of the geometric properties of circular DNA by intercalation of ethidium bromide. The intercalation of ethidium bromide into negative supercoiled DNA is energetically preferred compared to intercalation into the relaxed form of DNA as positive supercoils have to be introduced.

visible in the presence of unbound ethidium bromide. Ethidium bromide can be added to the gel and running buffer during electrophoresis, making post-staining of the gel unnecessary. The fluorescent ethidium cation migrates to the cathode during electrophoresis.

When performing longer electrophoresis runs, the running buffer should contain ethidium bromide since smaller, faster migrating fragments will otherwise be stained only weakly. For certain applications (e.g., blotting of the gel), the gels are stained after the electrophoresis. The blotting efficiency of RNA is diminished in the presence of ethidium bromide. Lanes for staining are cut from the gel and stained separately to control quality and size standard.

In agarose gels, approximately 10–20 ng of double-stranded DNA can still be detected. DNA with intercalated dye has a reduced mobility in the gel (approx. 15%). Since ethidium bromide intercalates into the DNA; it is a strong mutagen and should be handled with extreme care.

Influence on DNA Geometry Ethidium bromide changes the superhelical density of circular DNA molecules (form I) through a reduction of the negative supercoiling. Topoisomers with negative superhelical density turn into the relaxed form (increase of entropy). This conversion is favored over intercalation of ethidium bromide to give linear DNA fragments for thermodynamic reasons. Further intercalation of ethidium bromide will induce positive supercoiling. This process is less favorable compared to linear DNA (decrease of entropy, Figure 27.20).

For CsCl gradient density centrifugation the ethidium bromide concentration should be saturated. All supercoiled DNA is transformed into the conformation with positive supercoiling with a lower amount of ethidium bromide than for the relaxed forms. The positive supercoiled DNA has a lower density than linear (chromosomal) DNA and can be separated from the chromosomal contamination. The intercalation of ethidium bromide can also be used to analyze the conformation of circular DNA.

Other Fluorescent Dyes In recent years various intercalating dyes on the basis of asymmetric cyanine substance classes have been developed. These dyes are highly sensitive and less

mutagenic than ethidium bromide. A commonly used dye is SYBR®Green with an excitation maximum of 492 nm and a second absorption maximum at 284 nm. The emission maximum is at 519 nm. The dye can be used in fluorescence reader instruments, allowing exact quantification of the nucleic acid solution by comparison with standards.

Newly developed variants of this substance class display a higher affinity to single-stranded DNA or RNA (e.g., SYBR®Green II, OliGreen®). The dyes do not specifically bind to single-stranded nucleic acids but the quantum yield is drastically increased when interacting with single-stranded RNA. Other dyes, like TOTO-1 and YOYO-1, bind with much higher affinity to DNA than ethidium bromide. This can be used for certain applications when higher sensitivity is a decisive factor. Some of the new dyes (e.g., TOTO-3, YOYO-1, JOJO-1) cannot be excited with UV light and can only be used with laser induced fluorescence (LIF).

27.3.2 Silver Staining

Silver stains can only be performed with polyacrylamide gels as agarose gels yield too high a background staining. The gels need to be poured using high quality reagents and should be handled with extreme care as all protein or nucleic acid contamination will result in high background staining.

Silver staining is a less commonly applied method for the detection of nucleic acids. The advantage of this method, as for proteins, is its sensitivity. Very small amounts of nucleic acids (up to 0.03 ng mm^{-2}) can be detected. There is no need for mutagenic or radioactive detection reagents. The method is time consuming and background staining can be high.

Silver staining is based on the change of redox potential in the presence of nucleic acids (or proteins). The reduction of silver nitrate to silver is catalyzed. The metallic silver precipitates on the nucleic acids if the redox potential is higher than that in the surrounding solution. These conditions can be achieved through the choice of buffer and reagents. Recent analysis found that the purine bases account for this reaction.

27.4 Nucleic Acid Blotting

27.4.1 Nucleic Acid Blotting Methods

Electroblotting, Section 11.7

For further analysis of the nucleic acids they are separated by gel electrophoresis or transferred to a membrane. The fixed nucleic acids can be identified and analyzed by hybridization with labeled probes of known sequence. The nucleic acids can be transferred to a membrane by various methods: capillary blotting, vacuum blotting, and electroblotting. Figure 27.21 displays the principle underlying each method.

While blotting and subsequent hybridization with certain labeled probes was once the first choice to analyze nucleic acids (e.g., to detect chromosomal rearrangements, mutations, etc.), newer techniques (e.g., PCR, next generation sequencing) are nowadays applied more frequently than nucleic acid blotting, although blotting remains a useful, fast, and inexpensive way for nucleic acid analysis.

Capillary blotting can be performed with the least technical effort. The nucleic acids are transferred by capillary forces to the membrane using paper towels, through which the blotting buffer is soaked through the gel and membrane.

Vacuum blotting is performed using a vacuum chamber with the membrane attached and the nucleic acids are transferred through the gel. Capillary and vacuum blotting can only be performed with agarose gels. The nucleic acids will not move from polyacrylamide gels due to the lower pore size. For polyacrylamide gels, electroblotting systems are used. The nucleic acids are transferred to the membrane using an electric field. Electroblotting can be performed using a tank filled with buffer or by semi-dry blotting where the membrane and the filter are in contact with wet filter paper.

27.4.2 Choice of Membrane

For nucleic acid blotting two types of membranes are used: nitrocellulose and nylon membranes. Nitrocellulose has long been used but is increasingly being replaced by nylon (or poly

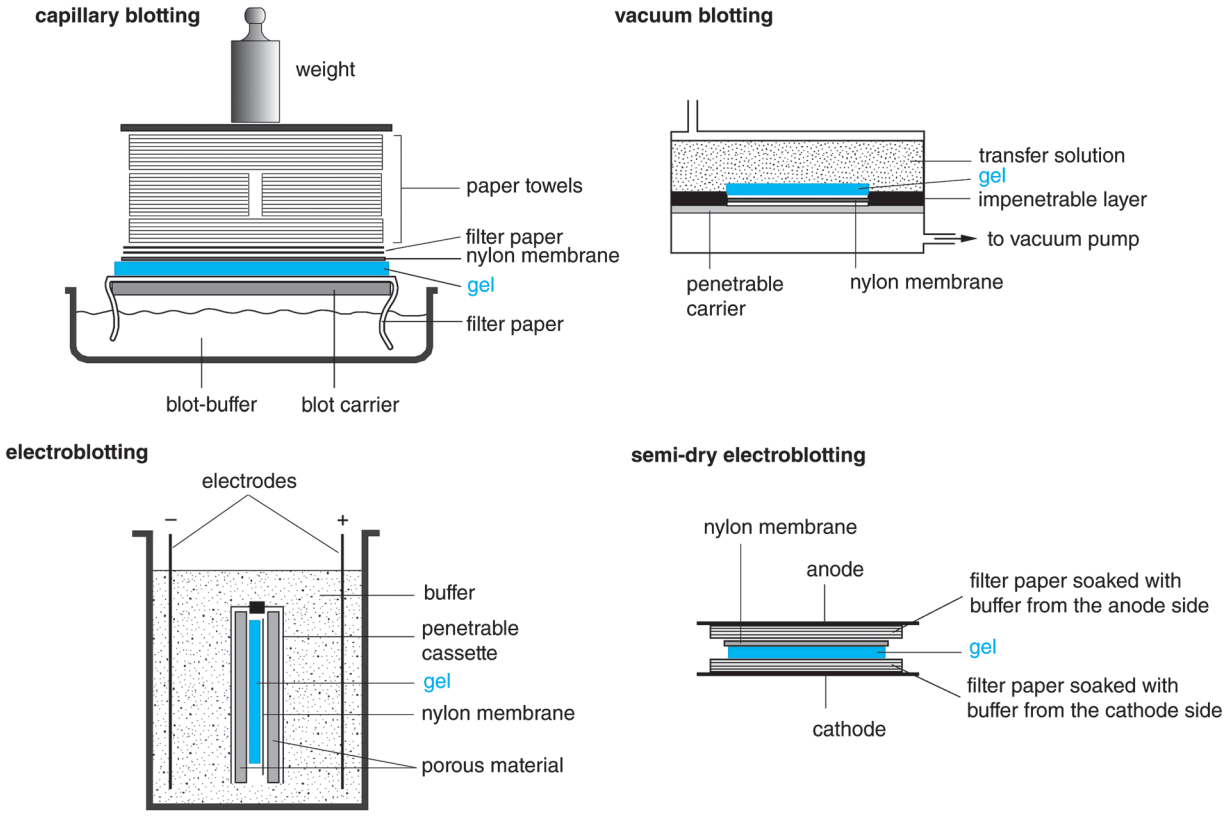


Figure 27.21 Schematic drawing of the different blotting techniques.

(vinylidene fluoride), PVDF) membranes with improved handling and binding properties. Table 27.9 gives an overview of membranes commonly used and their properties.

The nucleic acids are bound covalently to the surface of the nylon membrane, fixing them better to the material. The filters can be used several times. The binding to nitrocellulose is non-covalent.

The advantages of nylon membranes are manifold: higher stability, higher binding capacity, and better fixation. Nitrocellulose is harder to handle and is more fragile.

Nylon and PVDF membranes can yield a strong background signal, which can be reduced by the choice of suitable blocking reagents. If the membrane needs to be hybridized more than once, nylon membranes are the material of choice.

27.4.3 Southern Blotting

In 1975, E. Southern was the first to immobilize DNA separated by gel electrophoresis to a nitrocellulose membrane. Since then, the transfer of DNA from gels to a membrane has been called Southern blotting.

Table 27.9 Properties of different blotting membranes. Source: according to Ausubel, F.M., Brent, R.E., Moore, D.D., Smith, J.A., Seidman, and J.G., Struhl, K. (1987–2005) *Current Protocols in Molecular Biology*, John Wiley & Sons, Inc., New York.

Property	Nitrocellulose	Improved nitrocellulose	Neutral nylon membrane	Positively charged nylon membrane
Application	ssDNA, RNA, proteins	ssDNA, RNA, proteins	ssDNA, dsDNA, RNA, proteins	ssDNA, dsDNA, RNA, proteins
Binding capacity (μg nucleic acid cm^{-2})	80–100	80–100	400–600	400–600
Type of nucleic acid binding	Non-covalent	Non-covalent	Covalent	Covalent
Size restrictions for transfer	500 nt	500 nt	50 nt or bp	50 nt or bp
Resilience	Bad	Good	Good	Good
Multiple hybridizations	Bad (become brittle)	Bad (loss of signal intensity)	Good	Good

Southern blotting can be described by three main steps: preparation of the gel, transfer, and immobilization of the DNA on the membrane.

The efficiency of the transfer can be strongly improved by partial depurination of the DNA in the gel. The gel is treated with dilute hydrochloric acid, whereby purines are separated from the DNA backbone. When the DNA is denatured, either within the gel or during blotting, the phosphodiester bonds on the apurinic nucleotides are broken leading to a fragmentation of the DNA during transfer. The procedure is necessary especially for large DNA fragments (>10 kB). However, the fragments should not be too small because small fragments are not fixed efficiently to the membrane. The denaturing step is dependent on the type of nucleic acid to be transferred. For genomic DNA denaturing is essential. Usually, the gels are not stained with ethidium bromide to avoid gel and separation artifacts.

If nitrocellulose is used for blotting, the transfer buffer has to be high in salt for efficient binding to the membrane. Usually, 20 × SSC buffer is used containing sodium chloride and sodium citrate. If a denaturing step was performed, the gel needs to be neutralized before blotting as DNA does not bind to nitrocellulose at above pH 9.

For nylon membranes, the DNA can also be transferred using 20 × SSC buffer (with a previous denaturing step) or the denaturing step can be performed during blotting using an alkaline transfer buffer (e.g., 0.4 M sodium hydroxide) and sodium chloride.

Capillary blots are usually transferred overnight whereas vacuum blots are performed within 1–2 h. The DNA then needs to be immobilized on the membrane. The crosslinking of the DNA to the membrane can be performed using UV light. The thymidine bases are covalently linked to the amino groups of the nylon membrane. The duration and strength of the UV crosslinking needs to be optimized as crosslinking that is too strong makes most of the thymidine bases unavailable for hybridization whereas weak crosslinking leads to loss of signal intensity. If the transfer is performed using alkaline buffer, immobilization of the DNA to the nylon membrane is not necessary. With nitrocellulose membranes, the DNA is non-covalently bound by incubation of blotted membranes by temperature (not higher than 80 °C and not longer than 2 h, as the nitrocellulose can ignite). For the transfer of DNA out of polyacrylamide gels, only electroblotting can be performed using nylon membranes.

Southern blotting can be used for genomic analysis where the genomic DNA is cut with different restriction enzymes, separated, and blotted. The DNA can be analyzed for a specific hybridization pattern using known gene probes. It can also be used to detect gene families or single-copy genes. With increasing knowledge of the human genome by deep sequencing, genomic analysis by Southern blotting is becoming less common. Southern blotting techniques are and have been used to detect similarities between different species ("Zoo blot"). The genomic DNA of different species (or strain subtypes) is cut with a restriction enzyme and hybridized with the probe of interest.

27.4.4 Northern Blotting

According the nomenclature of Southern blotting for the transfer of DNA to membranes, the transfer of RNA to membranes is termed Northern blotting (and the transfer of proteins is called Western blotting). Like electrophoresis methods, blotting techniques need to account for the different properties of DNA and RNA. Since RNA is often separated in denaturing gels, the denaturing step preceding the blotting is not necessary. However, the denaturing reagents used during electrophoresis need to be removed by soaking of the gel in dilute sodium hydroxide solution or by incubation of the filter at higher temperatures. If high molecular weight RNA molecules need to be transferred, the gel is incubated briefly in sodium hydroxide (0.05 M) to partially hydrolyze the RNA for easier transfer.

Nitrocellulose membranes are more often used for Northern blotting than for Southern blotting. RNA gels are usually blotted in 10× or 20 × SSC buffer. An alkaline transfer is possible but only with a very low concentration of sodium hydroxide (7.5 mM). RNA gels are blotted by vacuum or capillary blotting, but the transfer requires more time (usually two days for a capillary blot). The RNA is immobilized similarly to DNA on the membrane.

Northern blots are used to study the expression of certain genes in different cells or tissue using total RNA or poly-adenylated mRNA purified by oligo d(T) affinity chromatography. With low expression genes, the use of purified poly(A) mRNA is mandatory. To detect

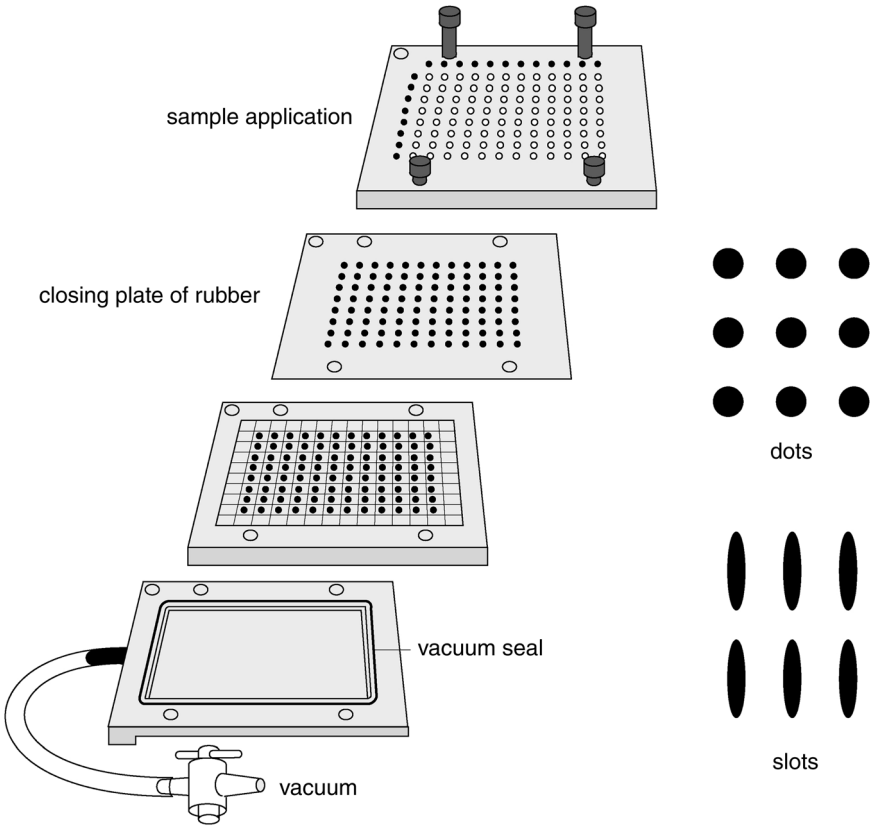


Figure 27.22 Scheme of the setup of a slot and dot blotting unit.

differences in the expression levels of mRNA of a certain gene, it is necessary to load equal amounts of RNA to the gel. To control the amount of RNA loaded and blotted to the gel, the blot is hybridized again with a probe for a gene that is equally expressed in all tissue or cells (housekeeping genes, usually glucose-6-phosphate dehydrogenase, β-tubulin, or β-actin mRNA). The signal strength obtained with the housekeeping probe should be similar for each of the different samples. A similar analysis is performed with RT-PCR (Chapter 29) and is usually the first choice method as Northern blots are not as quantitative as RT-PCR and require more hands-on time.

27.4.5 Dot- and Slot-Blotting

Dot- and slot-blotting are simple applications of membrane hybridization. The nucleic acids to be analyzed are blotted to the filter without prior separation by electrophoresis. The transfer is performed in dot-blotting units (Figure 27.22). With this set-up a large number of samples can be analyzed simultaneously. Dot- and slot-blots are used to analyze a large sample number for the presence or absence of a certain nucleic acid sequence.

27.4.6 Colony and Plaque Hybridization

A variant of the blotting technique is the generation of so-called colony or plaque filters for screening of cosmid or phage libraries. Colonies of bacteria grown on agar plates (colony hybridization) or phages (plaque filters) are transferred to membranes for hybridization with a certain probe. A huge number of colonies or phages can be screened for the presence of the sequence of interest. For colony or plaque hybridization, the stable nylon membranes are used. The membranes are transferred to the colony plate and colonies are lifted to produce exact copies of the agar plate. The direction of plate and filter need to be marked exactly to ensure that the colonies on the membrane can be allocated later to the colonies on the plate. Bacteria and phages are lysed using a denaturing solution (containing sodium hydroxide and sodium

chloride). The DNA is fixed to the membrane and the remaining, contaminating RNA hydrolyzed. Filters are neutralized and treated with UV light. To avoid hybridization artifacts, each agar plate is lifted twice and only those hybridization signals that can be detected on both membranes are considered positive.

27.5 Isolation of Nucleic Acid Fragments

To isolate pure DNA fragments, the fragments need to be separated with high resolution. This can be achieved by optimization of agarose concentration and electrophoresis conditions. The gel should not be overloaded as resolution will decrease.

Chaotropic agents Certain ions can disrupt hydrogen bonding and destroy chemical structures. These ions are negatively charged with large diameters and low charge density. Examples are I^-, ClO_4^-, and SCN^-.

DNA fragments of a defined sequence are the basis for various methods. Isolation methods for DNA fragments depend on the type of application they are needed for. Isolation of DNA from polyacrylamide gels will yield very pure material but the separation range and isolation efficiency are lower than for agarose gels. Isolation of fragments from agarose gels can result in contamination with polysaccharides, depending on the quality of the agarose. Very large DNA fragments (>5 kb) as well as small amounts of DNA are only isolated from an agarose gel with low efficiency. For preparative approaches (e.g., isolation of plasmid vectors), the DNA digested with restriction enzymes is loaded to several gel pockets. After separation the DNA fragment is cut out from the gel using a scalpel. The fragments are detected by ethidium bromide staining and UV light. For isolation of the fragments it is important to use long wavelength UV light and to keep the exposure time as short as possible to avoid damage of the DNA and crosslinking to the gel matrix.

The described methods can also be used to isolate or purify fragments that have not been separated previously by slab gels.

27.5.1 Purification using Glass Beads

Most commercially available DNA fragment isolation kits use glass beads. DNA can be bound to the glass surface in the presence of chaotropic salts (e.g., lithium acetate or sodium perchlorate). Hydrogen bridges within the agarose polymer are destroyed by high concentrations of chaotropic salts and the gel matrix is dissolved. At these salt concentrations, the DNA will adsorb to the silica surface of the glass beads. The adsorption is strongly pH dependent. Some protocols use pH indicators to ensure the correct pH (must be below 7.5 for optimal binding). The glass beads are either centrifuged or handled in columns. The adsorbed DNA is eluted from the glass beads after several washing steps to remove residual agarose and salt using low salt buffer at high pH (TE usually works well).

With this method, fragments with sizes between 40 bp and 50 kB can be isolated from 0.2–2% agarose gels. Higher molecular weight fragments (>4 kB) need more time and higher temperatures for isolation from the glass beads. The fragment yield will depend on the size of the DNA fragment. If the fragments are isolated from TBE gels, monosaccharides need to be present to chelate the borate anions. The method can also be used with some modifications for the isolation of DNA from polyacrylamide gels.

27.5.2 Purification using Gel Filtration or Reversed Phase

The principle of gel filtration has been discussed in Chapter 26.1.2. For fragment isolation this method is used to separate (radioactive labeled) nucleotides from reactions or to desalt the nucleic acid solutions. The choice of column material is dependent on the fragment size. The method is also frequently used for the purification of PCR fragments, to remove primers, nucleotides, and fluorescent dyes. Gel filtration methods are available for 96-well format.

27.5.3 Purification using Electroelution

This method is based, similarly to electrophoresis, on the migration of nucleic acids within the electric field. The method requires a higher instrumental effort and is less frequently used than commercially available kits, but if an instrument is available electroelution is cheaper than gel filtration columns.

With the simplest experimental set up, the agarose gel is captured in a dialysis tube containing electrophoresis buffer and submitted to an electric field. The DNA will migrate

according to the field out from the agarose piece and is trapped in the dialysis tube. Dependent on the type of electroelution instrument, the technical set up can vary; however, the DNA will always concentrate in the direction of the anode.

27.5.4 Other Methods

Oligonucleotides can be purified using denaturing polyacrylamide gels (Section 27.2.1). It is possible to separate *n*-mers from failure sequences $(n-1)$ or $(n-2)$. The oligonucleotide band can be visualized using fluorescence quenching. Efficient elution of the cut bands can be achieved by incubation of the small cut polyacrylamide gel pieces in sodium acetate solution. The oligonucleotides will diffuse into the solution.

If DNA fragments of low purity and yield are needed (e.g., for simple cloning experiments), agarose gel pieces can be centrifuged through silanized glass wool. The agarose matrix will be disrupted by the centrifugal forces and held back by the glass wool. Simple centrifugation of the gel pieces in a tube can also yield enough DNA fragment in the supernatant above the agarose pellet. These methods are not suitable if high purity and high yield fragments are needed.

27.6 LC-MS of Oligonucleotides

Markus Weber[1] and Eugen Uhlmann[2]

[1] *Currenta GmbH & Co. OHG, Chempark Q18, 51368 Leverkusen, Germany*
[2] *iNA ImmunoPharm GmbH, Zentastrasse 1, 07379 Greiz, Germany*

27.6.1 Principles of the Synthesis of Oligonucleotides

Synthetic oligonucleotides and their derivatives are important tools in molecular biology and in the development of new types of drugs, in particular antisense oligonucleotides, siRNAs, aptamers, antagomirs, and CpG adjuvants. These days their synthesis takes place on gram and kilogram scales, primarily by the phosphoramidite method on a solid phase (Figure 27.23). The step-by-step, computer-controlled synthesis takes place in the 3′ to 5′ direction. The first nucleoside residue is bound to the solid phase support (organic polymer or controlled pore glass) by its 3′-hydroxyl group and a base labile succinic acid. Orthogonal protective groups, like the acid labile 5′-*O*-dimethoxytrityl (DMT) protective group and base labile protective

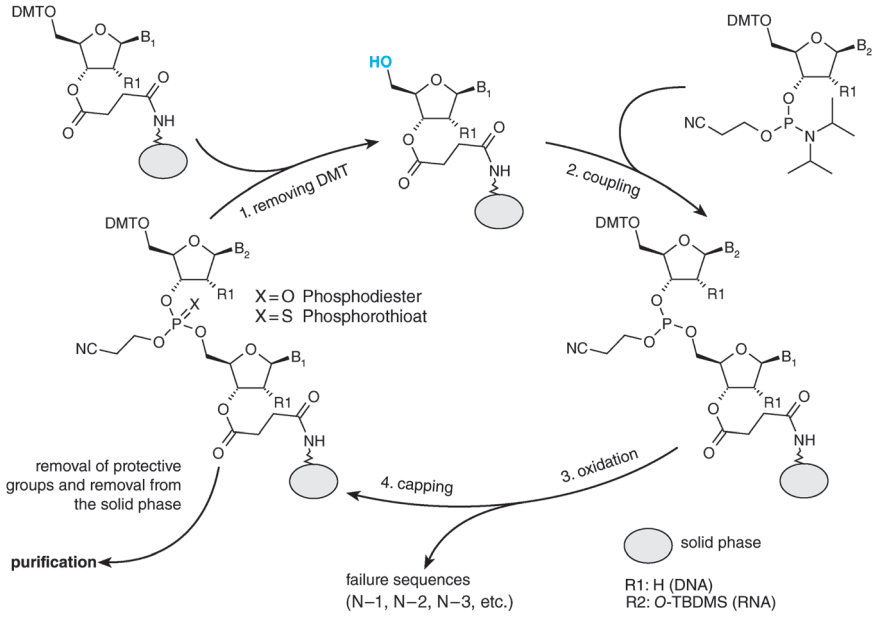


Figure 27.23 Schematic representation of the reaction cycle of oligonucleotide synthesis according to the phosphoramidite method on a solid phase (TBDMS: *t*-butyldimethylsilyl, DMT: dimethoxytrityl).

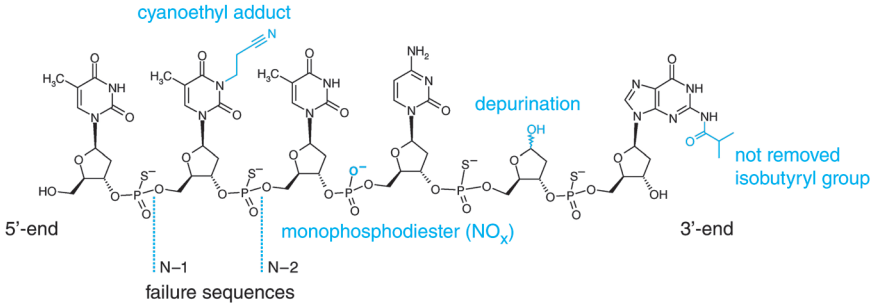


Figure 27.24 Possible by-products of the oligonucleotide (phosphorthioate) synthesis.

Protective groups Temporarily introduced groups that enable a molecule with several reactive functions to selectively react at only one of these functions.

Diastereomers Substances with several chiral centers that are not mirror images of one another and are therefore different compounds with different physical properties. For example, the phosphorothioate modified oligonucleotides have, besides the chiral β-D-deoxyribose, a further chiral center on the phosphate.

groups on the nucleobases and on the phosphate residue, allow the targeted exposure of reactive functions. In the first step, the DMT group is removed by treatment with diluted tri- or dichloroacetic acid. The free hydroxyl group is converted into a trivalent phosphite ester in a condensation reaction with 5′-O-DMT-nucleoside-3′-phosphoramidite, catalyzed by tetrazole. These esters can then be oxidized to a phosphotriester with iodine or to a thiophosphotriester with a sulfurizing agent like the Beaucage reagent. After complete chain synthesis and removal of the protective groups, the oligonucleotides can be phosphodiesters, phosphorothioates, or mixed backbone analogs, depending on the reagents employed. Since the substitution of an oxygen atom with a sulfur atom creates a chiral phosphate, in the course of making phosphorothioates a mixture of 2^n diastereomers results, where n is the number of internucleotide bonds. For example, a 20-mer oligonucleotide with 19 phosphorothioate modifications on the internucleotide bonds consists of 524 288 diastereomers. An acylation capping reaction is used to prevent excess 5′-hydroxy components from reacting in subsequent cycles of the coupling reaction. After multiple repetitions of the reaction cycle, corresponding to the length and composition of the desired sequence, the oligonucleotide is removed from the solid-phase column with concentrated ammonia and the protective groups are removed. RNA synthesis differs from DNA synthesis only in as far as an additional protective group is required for the 2′-hydroxyl group. Often, the *t*-butyldimethylsilyl (TBDMS) protective group is used, which is stable during the synthesis and can be removed in the very last step with triethylammonium fluoride.

Although the cycles of the phosphoramidite method operate with a very high yield of 98–99%, by-products can be present that are either the product of failed reactions during chain assembly or result from the final reactions to remove the protective groups (Figure 27.24). Since the coupling reactions do not operate with 100% efficiency, not only are oligonucleotides of the full length (N) present but also those of shortened lengths ($N-1$, $N-2$, $N-3$, etc.), which are missing one or more nucleotides. Interestingly, the reactions can also result in a nucleotide of greater than the expected length ($N+1$). These arise from a double addition during the tetrazole-catalyzed coupling reaction as a result of a minor cleavage of the acid labile DMT group, either on the monomer or on the growing chain. This side reaction happens most frequently during the condensation of deoxyguanosine, whose 5′-DMT group is the most labile of the four bases, due to the slightly acidic nature of the tetrazole catalyzed coupling reaction. In the case of incomplete sulfurization during the synthesis of phosphorothioates, reaction products contain a phosphodiester bond (mono-phosphodiester) in addition to the expected phosphorothioate. Another side reaction during the synthesis of purine-containing sequences is depurination (Figure 27.25), which takes place in an acidic environment. This refers to the hydrolysis of the N-glycoside bond between the nucleobase and deoxyribose due to protonation of the purine base at the 7-position. Possible side reactions of the deprotection are either the incomplete removal of the protective groups, such as the isobutyryl protective group on the exocyclic amino function of guanine, or the production of acrylonitrile adducts. The latter can come about

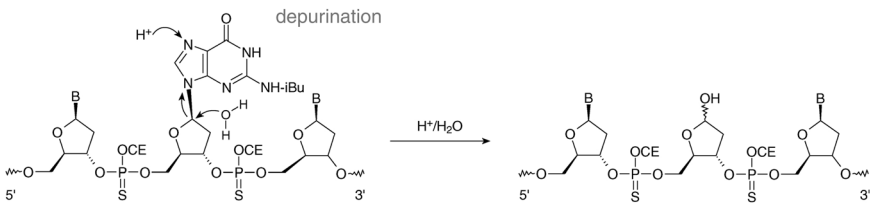


Figure 27.25 Depurination reaction (*i*Bu: isobutyryl).

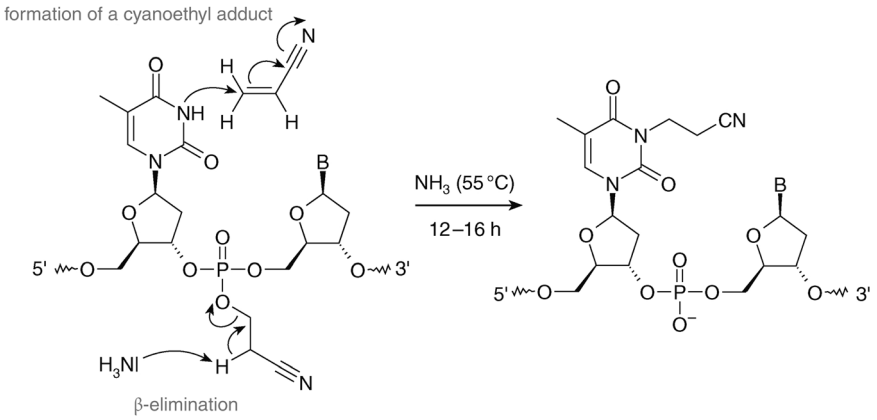


Figure 27.26 Formation of cyanoethyl adducts during the removal of the cyanoethyl protective groups with concentrated ammonia.

after β-elimination of the 2-cyanoethyl phosphate protective group and subsequent base-catalyzed addition of the acrylonitrile to the N3 of the thymine base (Figure 27.26). The by-products of oligonucleotide synthesis, due to their complexity, can only be partially separated by subsequent ion exchange or reversed-phase chromatography and are therefore only detectable by suitable analytical methods, such as the LC-MS method described in the following section.

27.6.2 Investigation of the Purity and Characterization of Oligonucleotides

While traditionally oligonucleotides with the natural phosphodiester internucleotide bond were of primary interest, more recently modified oligonucleotides now play a greater role, particularly in the area of therapeutic applications. The demands on the capabilities of the analytical methods have increased dramatically due to the increasing use of these modified oligonucleotides in recent years. Methods like polyacrylamide gel electrophoresis (PAGE), capillary gel electrophoresis (CGE), and anion exchange high performance liquid chromatography (HPLC) have been the mainstay of analytical techniques in the past, but the investigation of synthetic oligonucleotides, in particular, makes the use of HPLC-MS methods increasingly important. The online coupling of HPLC with a detection method based on mass spectrometry (MS) results in extremely powerful and conclusive HPLC-MS methods. A successful LC-MS analysis requires the best possible separation of the analytes with a HPLC method. Furthermore, it must be ensured that the HPLC method is compatible with the subsequent MS methods, which represents a significant hurdle in the development of new LC-MS techniques. Electrospray ionization mass spectrometry (ESI-MS) is the method of choice. It, however, requires the use of volatile buffer systems in the preceding HPLC. The direct infusion of the oligonucleotide to be investigated without prior HPLC purification is greatly complicated by the formation of cation adducts that are a product of the high affinity of oligonucleotides for sodium and potassium ions. Greig and Griffey have shown that the addition of strong bases like triethylamine (TEA) or piperidine strongly reduces the formation of adducts and thereby increases the sensitivity of ESI detection. For the investigation of complex mixtures by ESI-MS, a separation of the analytes by HPLC is essential. It turned out, however, that the mobile phases that led to a good separation of the analytes inhibited the ionization of the electrospray. As a result of the weakly hydrophobic character and the polyanionic nature of oligonucleotides they are ill suited to conventional reversed-phase (RP) HPLC. Ion pair reagents, which strengthen the interaction between the analytes and the stationary phase of the column, are therefore employed during the HPLC separation of oligonucleotides (Figure 27.27). Triethylammonium acetate (TEAA) and tetrabutylammonium bromide (TBAB) are two IP reagents that are frequently used during the separation of oligonucleotides by IP RP-HPLC. TBAB is, however, not volatile and can therefore not be used in combination with electrospray ionization. Although TEAA is a volatile ion pair reagent, it negatively impacts the sensitivity of the MS detection. The concentration of TEAA normally required for efficient separation leads, in general, to a significant loss of sensitivity of the MS detection. Apffel *et al.* were the first to use hexafluoroisopropanol/

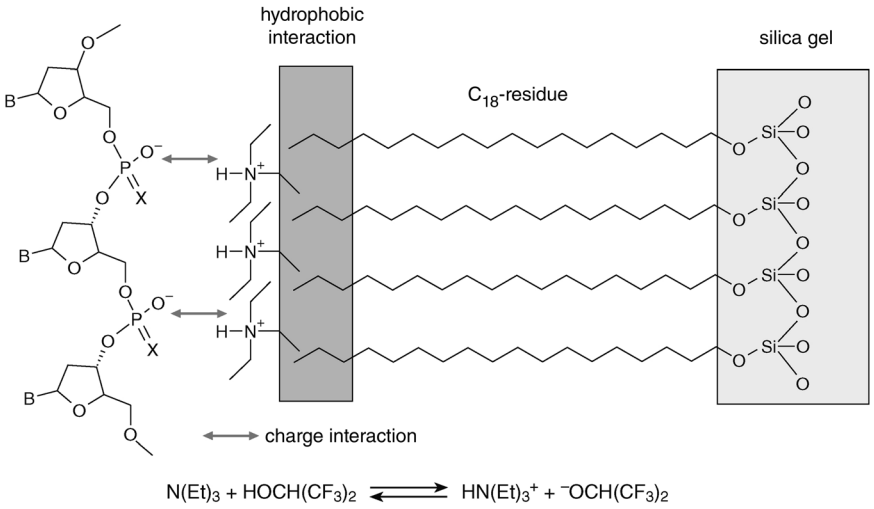


Figure 27.27 Schematic representation of the mechanism of separation of ion pair reversed-phase HPLC. Reversed phase chromatography means that the stationary phase is less polar than the eluent mixture. Typical stationary phases are porous silica gels, which have alkyl groups of various lengths bound to their surface. The chain length of the alkyl residues determines the hydrophobicity of the stationary phase. Most often C_{18} or C_8 alkyl chains are used.

triethylamine (HFIP/TEA) as an ion pair reagent and thereby achieved a high efficiency of the HPLC separation while maintaining the high sensitivity of MS detection and low adduct formation. Apffel attributed the increased MS sensitivity of the HFIP/TEA buffer to the different boiling points of HFIP, acetic acid, and TEA. Since acetic acid (boiling point 118 °C) has a higher boiling point than TEAA (89 °C), TEA evaporates preferentially, leading to a decrease in the pH of the HPLC eluent during the electrospray process. The drop in pH leads to protonation of the negatively charged oligonucleotide and therefore to a decrease in the sensitivity of the MS detection. During the desolvation of a HPLC eluent from HFIP/TEA buffer and analytes, in contrast, the HFIP preferentially evaporates, which leads to an increase of the pH value and therefore a deprotonation of the phosphate groups in the oligonucleotide backbone. The resulting negatively charged oligonucleotide can be evaporated into the gas phase during the ESI process and can be detected with high sensitivity. Gilar *et al*. further optimized the HFIP/TEA buffer system originally introduced by Apffel *et al*. for the separation of oligonucleotides by HPLC and thereby facilitated its broad adoption in the research and development of oligonucleotides.

Depending on the nature of the column and eluents, even nonpolar and polar analytes can be separated by RP-HPLC. If the polar characteristics are very pronounced, such as with oligonucleotides, a separation by normal RP methods is not possible. To increase the reactivity and thereby also the affinity of the polar substances to RP phases, ion pair reagents are generally used. These are characterized by undergoing a hydrophobic interaction with the RP phase and a charged interaction with the analyte. Ion pair reagents are used for the separation of oligonucleotides, which use alkyl residues to attract ammonium ions to increase the interaction between the analytes and the RP phase. Besides charged interactions, hydrophobic interactions between the RP phase and the hydrophobic bases of the oligonucleotides occur, which contribute to the total retention of the analytes. The separation efficiency of an ion pair method for the separation of oligonucleotides is primarily determined by the lipophilic nature of the ammonium cation. In addition, the counter-ion also has an impact on the separation. Gilar explains the high separation efficiency of the HFIP/TEA buffer as a result of the decreased solubility of protonated TEA molecules in HFIP, compared to acetic acid, which increases the surface concentration of the cation on the RP phase.

27.6.3 Mass Spectrometric Investigation of Oligonucleotides

HPLC systems coupled to mass selective detectors use electrospray ionization in which the analytes, dissolved in the separation buffer used to elute them from the HPLC column, are injected via a capillary into the ion source. Under normal atmospheric pressure, an electric field of several kilovolts is applied to the LC capillary as the ion source passes through it to form a

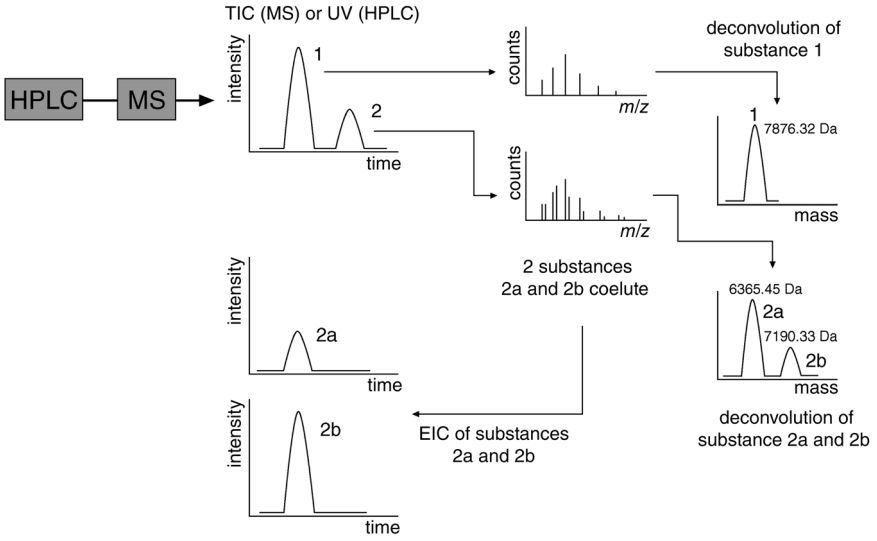


Figure 27.28 Schematic representation of the LC-MS analysis of a complex mixture of three components, of which two (2a and 2b) cannot be separated by chromatography.

finely dispersed spray of highly charged solvent droplets with a diameter of a few micrometers. The analysis of oligonucleotides, which due to their phosphate backbone form very slightly negatively charged molecular ions, takes place in negative ion mode, in which the LC capillary receives a positive charge. The ionization is particularly effective when the oligonucleotide is already in a deprotonated form due to the use of a suitable HPLC buffer system. This can be achieved by the use of buffers that have an alkaline pH value during the electrospray. The ions in the solvent droplets move into the gas phase through the process of desolvation, during which, dependent on the molecular weight of the oligonucleotide, primarily multiply charged ion molecules are formed. The charge distribution is determined mainly by the molecular weight of the analyte, but may also be influenced by the type of HPLC buffer, as well as the device parameters. A fragmentation of the analytes is not observed due to the low thermal load during the ionization as part of the electrospray procedure. After the transfer of the analytes into the gas phase, the mass analysis of the ions takes place in an ion mass spectrometer. For the analysis of oligonucleotides, HPLC coupling offers the advantage that complex substance mixtures can be investigated in a relatively simple manner. In addition, the chromatographic purification offers the possibility of an almost complete removal of salts, which would otherwise inhibit the electrospray process. By coupling HPLC with the ESI-MS one receives, in addition to the UV chromatogram, further chromatographic data, the so-called total ion current (TIC) chromatogram, which usually correlates well with the UV chromatogram detected at 260 nm. The mass spectrometric detection allows the visualization of a mass spectra at every point in the TIC chromatogram (Figure 27.28). The electrospray ionization of oligonucleotides generally leads to the formation of a series of multiply charged ions, which carry a variable number of negative charges in their backbones. Therefore, in the mass spectra of an oligonucleotide there are always a series of ion signals that differ by exactly one charge, z. As a result, for a molecule of mass m, a series of different values of m/z are always detected. Since ESI-MS does not detect the mass directly, but instead the ratio m/z, multiply ionized molecules with a relatively high mass can still be measured. The intensity of an ion signal always depends on the statistical probability that the corresponding ion is formed during the electrospray process. In the ideal case the intensity distribution forms a Gaussian distribution curve. The actual form of the intensity distribution and the position of the maximum are, however, dependent on the choice of MS parameters, since these can strongly influence the transmission of individual ions. The formation of ion series prevents the direct determination of the molecular weight from the mass spectra of the oligonucleotides. The molecular weight can be obtained, however, by deconvolution of the data using the measured m/z values of the individual charged states. The mass spectra recorded by an ion trap in full scan mode can be used to generate an extracted ion chromatogram (EIC). This involves using the individual m/z values of an ion series of a compound to calculate a chromatogram trace, which can be used to see at what time point a particular component is eluted. An EIC can be used to generate

Gauss (or normal) distribution – a symmetrical distribution in the form of a bell curve, with which many random processes in nature can be described.

chromatogram traces of co-eluting substances of differing molecular weights. These can then be used like any normal UV or TIC chromatogram to quantify substances; this means that even substances that cannot be separated by HPLC can be quantified, provided they are of differing molecular weights.

27.6.4 IP-RP-HPLC-MS Investigation of a Phosphorothioate Oligonucleotide

This section explains the investigation of a synthetic oligonucleotide using IP-RPHPLC-MS by using the example of a 20-mer phosphorothioate oligonucleotide. There are no other chemical modifications besides the phosphorothioate modification of the oligonucleotide backbone. Table 27.10 shows the sequence of the phosphorothioate.

The objective of IP-RP-HPLC-MS based LC-MS analysis is to identify by-products of the oligonucleotide synthesis based on their molecular weight and therefore to draw conclusions about the purity of the main product of the synthesis. The UV or TIC chromatograph shown in Figure 27.29 provides an overview of the number of contaminants present. In addition to the

Table 27.10 Identification of the by-products from the synthesis of the 20 mer phosphorothioate oligonucleotide based on their molecular weight. The by-products, which make up less than 0.5% of the total UV trace, are not discussed here.

ID	MW (Da)	AMW (Da)	$MW_{Ber.}$ (Da)	Identification[a)]	% UV
A	6665.6	0	6665.4	N	78.1
B	6532.7	−132.9	6532.2	*N*-Guanine + H_2O	
C	6649.8	−15.7	6649.3	N_{ox}	
D	6320.3	−345.3	6320.1	N − G	11.6
E	7010.8	+345.2	7010.6	N + G	5.3
F	6719.1	+53.5	6718.4	N + CE	1.8
G	5974.9	−690.7	5974.8	N − 2G	2.4
H	5629.6	−1036.0	5629.5	N − 3G	0.8

a) N_{ox}: Mono phosphodiester.
N-Guanine + H_2O: depurinated.
N + CE: cyanoethyl adduct.
Sequence of the desired oligonucleotide:
G∗G∗G∗G∗G∗A∗G∗C∗A∗T∗G∗C∗T∗G∗G∗G∗G∗G∗G∗G. where ∗ represents a phosphorothioate-internucleotide bond.

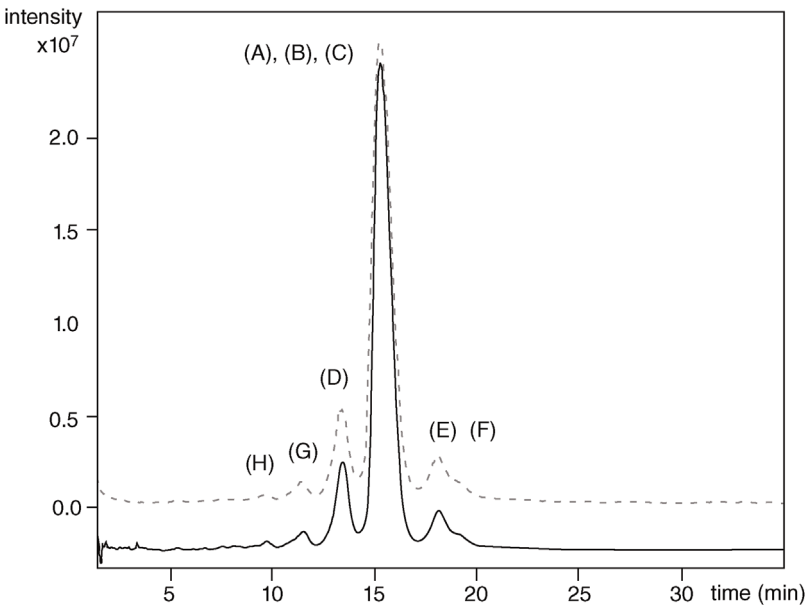


Figure 27.29 Ion pair RP-HPLC-separation of a 20-mer phosphorothioate-oligonucleotide. The total ion current (TIC, - - -) of the ESI detection corresponds to conventional UV (___) detection at 260 nm.

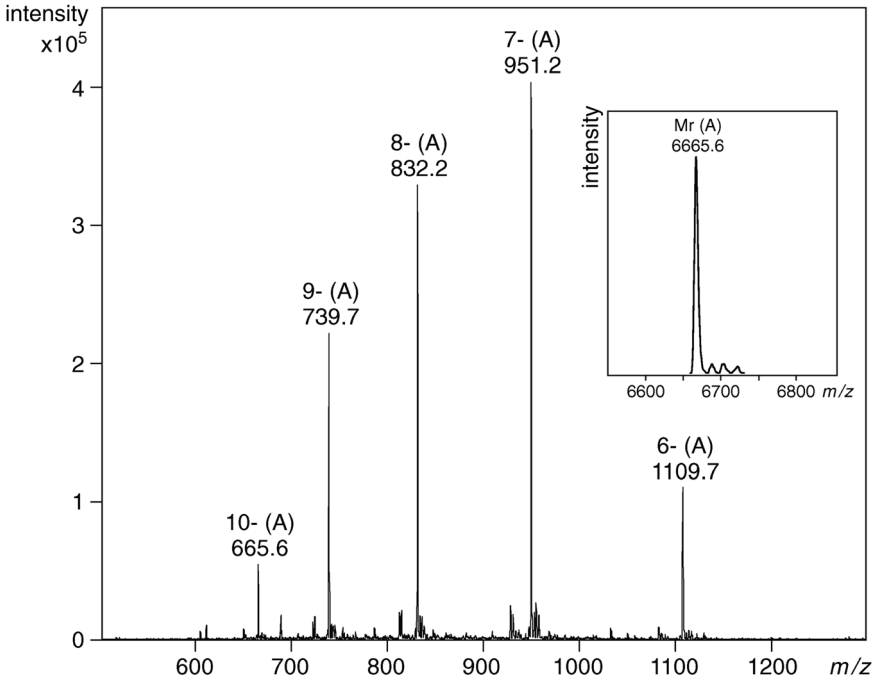


Figure 27.30 ESI spectra of the main components (A) of the chromatographic separation. The molecular weight of components (A) is determined by deconvolution of the ion series (A).

contaminants separated by HPLC, other contaminants that are not separable by chromatography can be present, which cannot be identified on the basis of the UV or TIC traces. Provided they are of differing molecular weight, they can be detected by ESI mass spectrometry and their content estimated via the EIC method described earlier.

With the aid of mass spectrometric detection, which is carried out in addition to conventional UV detection, it is possible to determine the molecular weight of compounds eluted from the HPLC into the mass spectrometer. Figure 27.30 shows a typical example of a measured MS spectrum of the desired main component (A). It shows a series of signals of the negatively charged ions typical of oligonucleotides. In this case, five signals were measured, which can be attributed to species with a differing number of negative charges (6–10) in the phosphorothioate backbone. Depending on the measurement conditions, a different number of charged states may be found in the primary spectrum. By deconvolution the molecular weight of the analyte can be calculated from the *m/z* values of an ion series. Here a molecular weight of 6665.6 Da was measured, which is sufficiently close to the calculated mass of the desired oligonucleotide of 6665.4 Da.

The molecular weight of two analytes (compare compounds E and F in Figure 27.29) incompletely separated by HPLC chromatography are easily determined based on their mass spectra (Figure 27.31). The series of the MS signals of the components of E correlate due to their higher signal intensity with the signal of the higher UV intensity of the double peak (E), (F) in the UV lane. Using the higher intensity signals of series E, a molecular weight of 7010.8 Da is calculated, which corresponds to an oligonucleotide extended by one nucleotide (N + G) (Table 27.10). The signals of the series F result in a mass of 6719.1 Da after deconvolution, which correspond to the cyanoethyl adduct of the oligonucleotide (N + CE). The signals E and F, which could not be completely separated chromatographically, can be unequivocally assigned to the two by-products N + G (E) and N + CE (F) (Table 27.10). On the basis of these simple considerations, it is possible to determine the molecular weight of components A and D to H without great difficulty (Figure 27.29). The by-products B and C of the synthesis cannot be separated from the main product A of the reaction by HPLC; this means that all three compounds appear in the UV or TIC detection traces as a single signal. In this case it is, however, possible to use the mass-dependent detection of these components to identify them. Just ahead of the main components in the mass spectra, beside the signal from the main component A, there are two more ion series: B and C (Figure 27.32). Deconvolution allows determination of the molecular weights of the by-products B and C, which coelute with the target compound A. In this case, besides the desired oligonucleotide N of mass 6665.4 Da, two

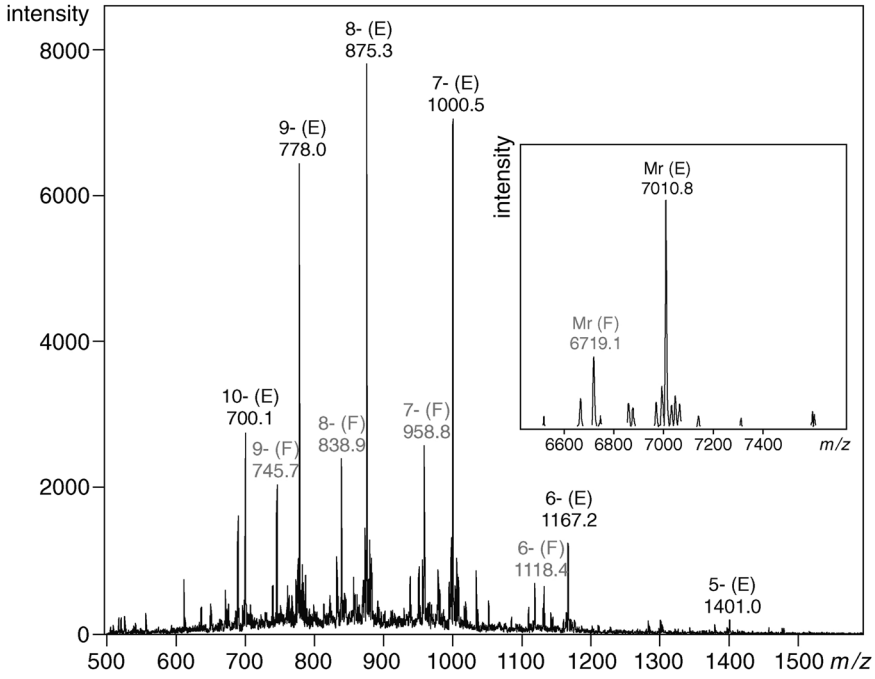


Figure 27.31 ESI spectra of the chromatographically incompletely separated components (E) and (F).

other by-products of masses 6532.7 Da (B) and 6649.8 Da (C) are detected, which correspond to the depurinated product (B) and the mono phosphodiester N_{ox} of the oligonucleotide (C).

The *m/z* value of the differentially charged analytes of the ion series A, B, and C can also be used to extract an ion chromatogram. The EIC method is particularly helpful in this case, since the by-products of the synthesis (ion series B and C) cannot be separated chromatographically from the main product A. In this manner the by-products B and C, which coelute with the main product, can be quantified (Figure 27.33).

The mass spectrometrically determined molecular weight of the main product A can be used to make a comparison with the expected molecular weight to confirm its identity. In addition, in many cases the difference in molecular weight between the main product and by-products can

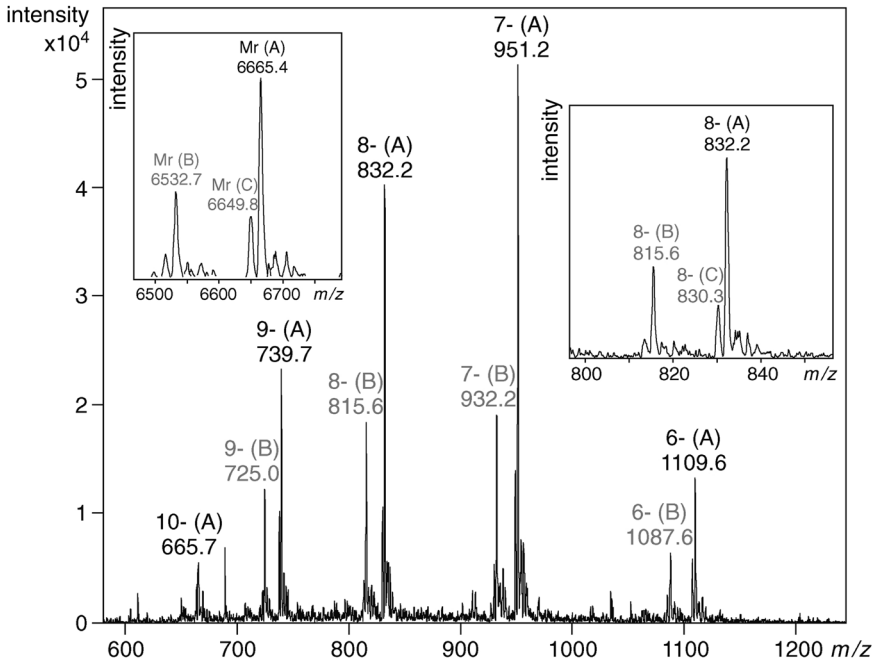


Figure 27.32 ESI spectra from the leading edge of the main components detect three components of differing molecular weight.

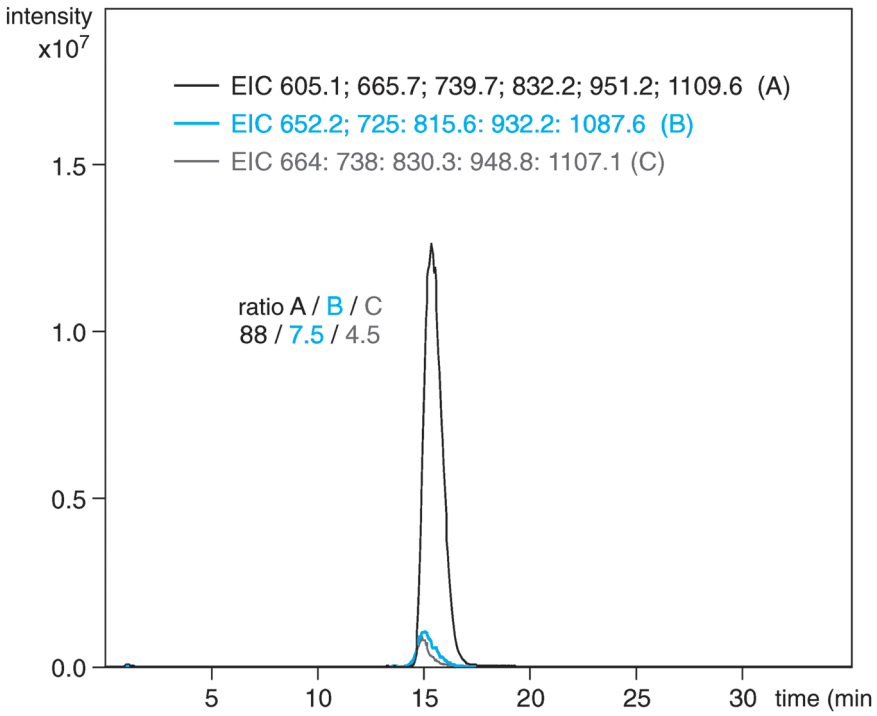


Figure 27.33 Extracted ion chromatograms (EIC) of the components A, B, and C. By integration of the chromatogram traces, the relationship of the co-eluting compounds A, B, and C to one another can be determined.

be used to determine the identity of the by-products. (Table 27.10). In this example, what appears in the UV-HPLC chromatogram as a single uniform peak (A, B, C) (Figure 27.29) is revealed by the EIC trace to be a mixture of the desired main product A (88%), as well as the depurinated product B (7.5%) and the mono phosphodiester C (4.5%) (Figure 27.33).

Further Reading

Section 27.1

Ausubel, F.M. Brent, R., Kingston, R.E., Moore, D.D., Smith, J.A., Seidman, J.G., and Struhl, K. (1987–2005) *Current Protocols in Molecular Biology*, John Wiley & Sons, Inc., New York.

Roberts, R.J. *et al*. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.*, **31**, 1805–1812.

Roberts, R.J., Vincze, T., Posfai, J., and Macelis, D. (2010) REBASE – a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **38**, D234–D236.

Section 27.2–27.5

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Smith, J.A., Seidman, J.G., and Struhl, K. (1987–2005) *Current Protocols in Molecular Biology*, John Wiley & Sons, Inc., New York.

Chrambach, A., Dunn, M.J., and Radola, B.J. (eds) (1987) *Advances in Electrophoresis, Volume 1*, VCH-Verlagsgesellschaft, Weinheim.

Darling, D.C. and Brickell, E.M. (1994) *Nucleinsäure-Blotting. Labor im Fokus*, Spektrum Akademischer Verlag, Heidelberg.

Glasel, J.A. and Deutscher, M.E. (1995) *Introduction to Biophysical Methods for Protein and Nucleic Acid Research*, Academic Press, New York.

Grossman, L. and Modave, K. (eds) (1980) *Nucleic Acids: Part I, Methods in Enzymology*, vol. 65, Academic Press, New York.

Hafez, M. and Hausner, G. (2012) Homing endonucleases: DNA scissors on a mission. *Genome*, **55**, 553–569.

Krieg, E.A. (ed.) (1996) *A Laboratory Guide to RNA: Isolation, Analysis and Synthesis*, Wiley-Liss, New York.

Martin, R. (1996) *Elektrophorese von Nucleinsäuren*, Spektrum Akademischer Verlag, Heidelberg.

Miller, J.M. (2013) Whole-genome mapping: a new paradigm in strain-typing technology. *J. Clin. Microbiol.*, **51** (4), 1066–1070.
Nassonova, E.S. (2008) Pulsed field gel electrophoresis: theory, instruments and application. *Cell Tissue Biol.*, **2** (6), 557–565.
Nowacka, M., Jockowiak, P., Rybarcyk, A., Magaz, T., Strozycki, P.M., Barciszewski, J., and Figlerowicz, M. (2012) 2D-PAGE as an effective method of RNA degradome analysis. *Mol. Biol. Rep.*, **39**, 139–146.
Rickwood, D. and Hares, B.D. (eds) (1990) *Gel Electrophoresis of Nucleic Acids: A Practical Approach*, IRL Press, Oxford.
Salieb-Beugelaar, G.B., Dorfman, K.D., van den Berg, A., Eijkel, J.C.T. (2009) Electrophoretic separation of DNA in gels and nanostructures. *Lab Chip*, **9**, 2508–2523.
Sambrook, J. and Russell, D.W. (2001) *Molecular Cloning: A laboratory Manual*, 3rd edn, Cold Spring Harbor Press, Cold Spring Harbor.

Section 27.6

Apffel, A., Chakel, J.A., Fischer, S., Lichtenwalter, K., and Hancock, W.S. (1997) New procedure for the use of high-performance liquid chromatography-electrospray ionization mass spectrometry for the analysis of nucleotides and oligonucleotides. *J. Chromatogr. A*, **777**, 3–21.
Engels, J.W. (2013) Gene silencing by chemical modified siRNAs. *New Biotechnol.*, **30** (3), 302–307.
Gilar, M. (2001) Analysis and purification of synthetic oligonucleotides by reversed-phase high-performance liquid chromatography with photodiode array and mass spectrometry detection. *Anal. Biochem.*, **298**, 196–206.
Gilar, M., Fountain, K.J., Budman, Y., Holyoke, J.L., Davoudi, H., and Gebler, J.C. (2003) Characterization of therapeutical oligonucleotides using liquid chromatography with on-line mass spectrometry detection. *Oligonucleotides*, **13**, 229–243.
Gilar, M., Fountain, K.J., Budman, Y., Neue, U.D., Yardley, K.R., Rainville, E.D., Russell, R.J. II, and Gebler, J.C. (2002) Ion-pair reversed-phase high-performance liquid chromatography analysis of oligonucleotides: retention prediction. *J. Chromatogr. A*, **958**, 167–182.
Greig, M. and Griffey, R.H. (1995) Utility of organic bases for improved electrospray mass spectrometry of oligonucleotides. *Rapid Commun. Mass Spectrom.*, **9** (1), 97–102.
Kusser, W. (2000) Chemically modified nucleic acid aptamers for *in vitro* selections: evolving evolution. *Rev. Mol. Biotechnol.*, **74**, 27–38.
Martin, R. (1996) *Elektrophorese von Nucleinsäuren*, Spektrum Akademischer Verlag, Heidelberg.
Matteucci, M.D. and Caruthers, M.H. (1981) Synthesis of deoxynucleotides on a polymer support. *J. Am. Chem. Soc.*, **103**, 3185–3191.
Uhlmann, E. (2000) Recent advances in the medicinal chemistry of antisense oligonucleotides. *Curr. Opin. Drug Discovery Dev.*, **3**, 203–213.
Uhlmann, E. and Vollmer, J. (2003) Recent advances in the development of immunostimulatory oligonucleotides. *Curr. Opin. Drug Discovery Dev.*, **6**, 204–217.
Warren, W.J. and Vella, G. (1995) Principles and methods for the analysis and purification of synthetic deoxyribonucleotides by high-performance liquid chromatography. *Mol. Biotechnol.*, **4**, 179–199.

28

Techniques for the Hybridization and Detection of Nucleic Acids

Christoph Kessler[1] **and Joachim W. Engels**[2]

[1]*PD Christoph Kessler, Consult GmbH, Icking, Schloßbergweg 11, 82057 Icking-Dorfen, Germany*
[2]*Goethe University Frankfurt, Institute of Organic Chemistry and Chemical Biology, Department of Biochemistry, Chemistry and Pharmacy, Max-von-Laue Straße 7, 60438 Frankfurt am Main, Germany*

The last two decades have seen the development of many new assay techniques for the detection and analysis of sequences of DNA or RNA. These highly specific and sensitive methods have become standard methods in molecular biology in a short span of time. Today they are used for:

- diagnosis of infective diseases: viral or bacterial identification;
- tissue and organ tolerance diagnostics: histocompatibility genes;
- cancer diagnosis and risk analysis: gene mutation analysis;
- diagnosis of inheritable diseases, pre-implantation diagnostics: gene and chromosome analysis;
- paternity testing, forensic medicine, animal breeding: DNA profiling;
- plant breeding: analysis of gene transfer, patterns of resistance;
- crop and wine analysis: tests for pathogens, resistance, or marker genes linked to new or modified gene products;
- molecular archeology and anthropology: gene analysis of mummies, archeological finds;
- production of recombinant human, pharmaceutically active proteins: quality control and specificity analysis;
- safety surveillance of genetic laboratories: contamination tests.

Other important fields of application are the elucidation of certain genetic changes, such as point mutations, deletions, and insertions, or triplet repeats, which cause disease in the fields of oncology, genetic disease, and chronic infection. Knowledge of the precise sequence changes responsible for disease is a prerequisite for the development of genetic diagnosis tests and gene therapy approaches. The Human Genome Project has provided a standard of reference for these efforts. Begun in the early 1990s and projected to last 15 years, the Human Genome Project elucidated completely the molecular structure and sequence of the human genome. The first sets of data of the complete human genome were published at the beginning of 2001 and updated by the Human Genome Consortium in October 2004. In 2008 the sequence of eight human genomes were published and at present efforts are underway to sequence 1000 genomes in the USA and 10 000 genomes in the UK (UK10K).

Disease can result from defined gene defects, chromosome aberrations, such as translocations, chromosome number abnormalities, or sub-chromosomal aberrations, such as amplifications or deletions. A well-known example is Down's syndrome, which is caused by the presence of three copies of chromosome 21.

Due to the great variety of potential gene changes, the methods used for the analysis of nucleic acids span a broad range. The analysis must be able to detect monogenetic mutations, those in single positions of the genome, as well as polygenetic mutation patterns, which involve a number of mutations. In many cases the mutations are polymorphic, meaning that more than one type of mutation is known, and new, spontaneous mutations may also arise. The type of

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

mutation determines the assay type: while defined mutations or simple mutation patterns are primarily detected with hybridization techniques, the analysis of variable, complex mutations, such as polymorphisms and spontaneous mutations, is increasingly performed with the new sequencing methods.

The sample type also influences the choice of assay type. Different techniques are required depending on whether the nucleic acids are isolated, amplified, or chromosome preparations, or whether they are fixed to solid membranes, in solution, or in cells, tissue, or organisms. An extreme case is *Drosophila* embryos fixed to glass. The analysis of sequence fragments or amplified DNA or RNA sequences involves *in vitro* nucleic acid analysis. *In situ* analyses are used in molecular cytogenetics to detect chromosome aberrations or to analyze endogenous or exogenous sequences in cells, tissues, or organisms. Different hybridization and sequencing methods are used in these techniques, depending on the nature of the changes in the nucleic acids. In this chapter we will discuss the most common methods of *in vitro* hybridization and detection. The nucleic acid sequencing will be explained in Chapter 30.

Fluorescent *in Situ* Hybridization in Molecular Cytogenetics, Chapter 35

DNA Sequencing, Chapter 30

28.1 Basic Principles of Hybridization

The complementary bases of DNA, A and T or C and G, are bound to one another via hydrogen bonds (Figure 28.1). Hydrogen bonds are non-covalent bonds of middle to low strength, depending on the hydrogen donor acceptor atom and the distance between them. This complementarity of base pairing is the basis of hybridization.

In 1961, Julius Marmur and Paul Doty discovered that the DNA double helix could be separated into two strands when denatured by heating it above its melting temperature. Allow a mix of single strands long enough to cool and they will hybridize back into double strands. Based on this observation, Sol Spiegelman developed the technique of nucleic acid hybridization. Spiegelman investigated whether newly synthesized bacterial mRNA is complementary to T2 DNA after an infection by T2 phages. In his experiment, he mixed ^{32}P-marked T2 mRNA and ^{3}H-marked T2 DNA. After denaturation of the double-stranded DNA into single strands and subsequent reassociation, he analyzed the nucleic acid mix by density gradient centrifugation. Since RNA has a higher density than DNA, it is possible to separate them in a caesium

Isolation of Plasmid DNA from Bacteria, Section 26.3.1

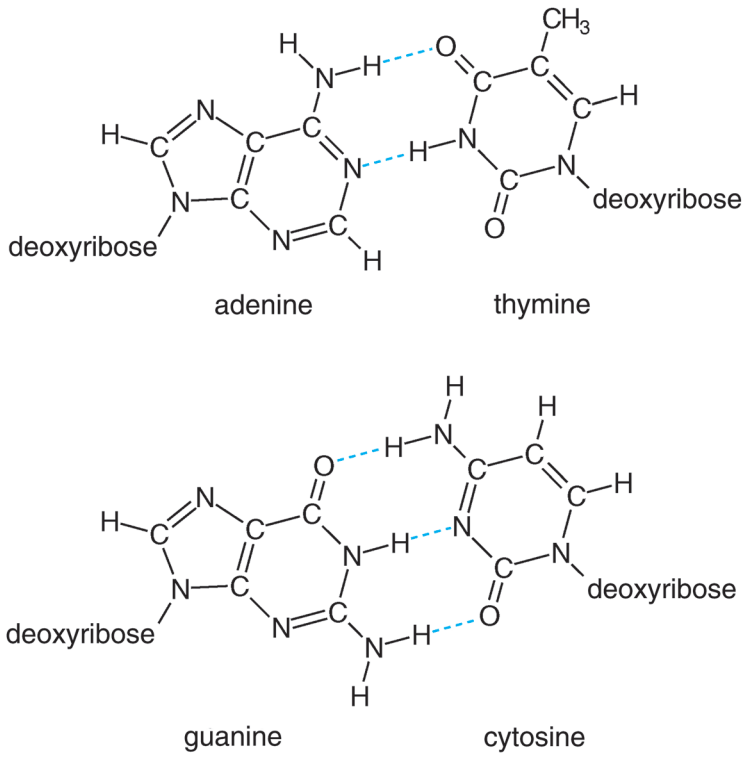


Figure 28.1 Complementarity of base pairing.

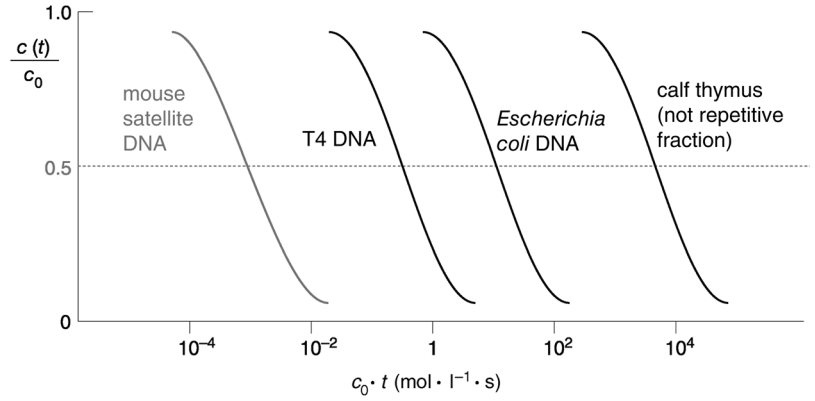


Figure 28.2 C_0t (cot) curves represent the renaturation of different thermally denatured DNA species. The left-hand curve, mouse satellite DNA, contains many repetitive sequences and therefore renatures quickly. Source: adapted according to Britten, R.J. and Kohne D.E. (1968) *Science*, **161,** 1530.

chloride (CsCl) density gradient. A measurement of the distribution of the radioactivity showed that a third band of RNA–DNA hybrids had appeared, in addition to the single-stranded RNA and the DNA double helices. In further experiments with T2 mRNA and DNA from unrelated organisms, no such hybrids were observed. This experiment showed that the correct sequence of complementary bases in the antiparallel nucleic acids is necessary for the hybridization reaction to take place.

Information about the complexity of a particular DNA sequence can be obtained from hybridization experiments. Sequences that appear frequently in the genome renature faster than those that only appear once. Eukaryotic DNA can be divided into four classes depending on its frequency within the genome: DNA molecules immediately renature to double helices when made up of inverse repeats (palindromes), which pair by folding back on themselves, and hairpin (or stem) loops. Highly repetitive sequences reform the helices somewhat slower. Then come the less repetitive sequences, and finally the unique sequences, which, under normal circumstances, are last to rehybridize.

The complexity of the DNA is expressed in the *cot value*: if c_0 is the concentration of single-stranded DNA at time point $t = 0$ and $c(t)$ the concentration of single-stranded DNA at time point t:

$$\frac{c(t)}{c_0} = \frac{1}{1 + kc_0t} \tag{28.1}$$

where k is the association constant, a kinetic constant. The function $c(t)/c_0$ is the proportion of double-stranded DNA at a particular time point, t (Figure 28.2). At a specific time, $t\phi$, 50% of the DNA strands are hybridized: $c(t)/c_0 = 0.5$; the value of $c_0 \times t\phi$ is the *cot value*.

Besides conclusions about the complexity of DNA, hybridization experiments can also be used to identify particular DNA sequences in mixtures. A known sequence, a probe, is labeled, either radioactively or non-radioactively, and is hybridized with the DNA to be analyzed. Identification results from the detection of the labeled hybrids.

28.1.1 Principle and Practice of Hybridization

One thing all hybridization techniques have in common is that the detection of the nucleic acid target molecules results from the sequence-specific binding of complementary labeled nucleic acid probes (Figure 28.3).

In general, the nucleic acid mixture to be analyzed is blotted onto a membrane or other solid support, such as onto the surface of a microtiter plate, or left in solution. Under carefully controlled, stringent conditions (Section 28.1.2), the nucleic acids are mixed with a solution containing the labeled probe and left to incubate at a fixed temperature. Labeled probes can be made from oligonucleotides, DNA fragments, PCR (polymerase chain reaction) products (amplicons), *in vitro* RNA transcripts, or artificial probes like PNA (Section 28.2.3).

Single-stranded RNA and DNA hybridize with one another, such that all three possible double helices are formed: DNA:DNA, DNA:RNA, and RNA:RNA. The probe hybridizes with the complementary target sequence. After completing the incubation, stringent washes are carried out

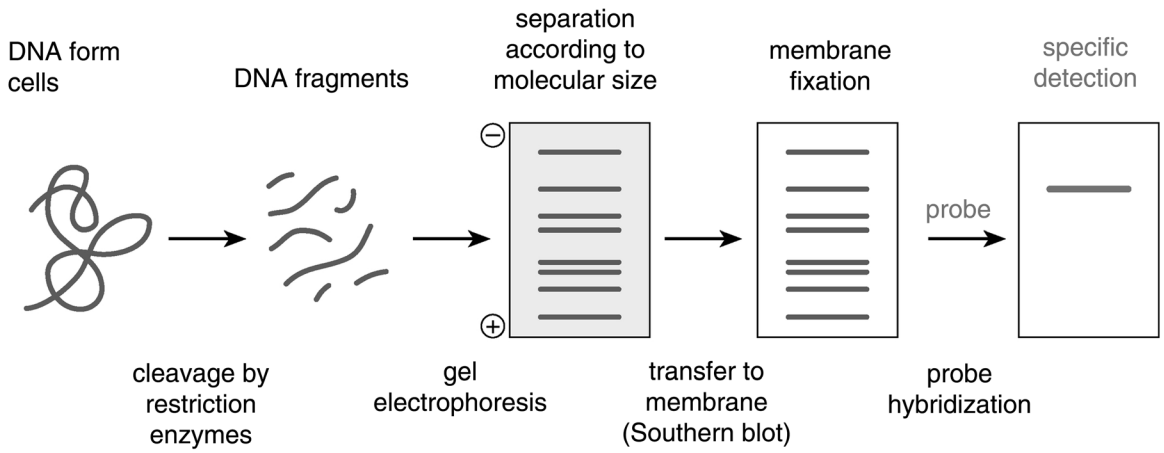


Figure 28.3 Nucleic acid detection by hybridization. Southern blots (named after Edwin Southern) are often used for this process. The polynucleotides, such as DNA fragments or PCR products, are incubated in a plastic bag for several hours with labeled probes at 60–70 °C. This temperature allows the DNA probes to hybridize with their complementary sequences on the blot. After washing, the position of the DNA segments complementary to the labeled probe is revealed by detection of the signal from the label.

to wash away unspecifically adsorbed probes. Techniques without wash steps, called homogeneous assays, have also been described. The sought target sequence is identified by measurement of the specific binding of the labeled probe; this specific binding is visualized by autoradiography or non-radioactive methods, which will be described in Sections 28.2 and 28.3.

28.1.2 Specificity of the Hybridization and Stringency

The specificity of hybridization is dependent on the stability of the hybrid complex formed, as well as the stringency of the reaction conditions. The stability of the hybrid correlates directly with its melting point (T_m). The T_m value is defined as the temperature at which half of the hybrids have dissociated. It is dependent on the length and base composition of the hybridizing section of sequence, the salt concentration of the medium, and the presence or absence of formamide or other helix-destabilizing agents, as well as the type of hybridizing nucleic acid strands (DNA:DNA, DNA:RNA, and RNA:RNA). For DNA:DNA hybrids the following formula applies to a first approximation:

$$T_m = 81.5\,°C + 16.6\log\left[c(Na^+)\right] + 0.41(\%\,G + C) - \frac{500}{n} \qquad (28.2)$$

In this equation, c (Na^+) is the concentration of Na^+ ions and n the length of the hybridizing section of sequence in base pairs. Because G/C hybrids contain three hydrogen bonds instead of two, like A/T hybrids, they increase the melting temperature more, which needs to be taken into account. The expression $500/n$ does not apply for a sequence longer than 100 bp.

The melting point of DNA:RNA hybrids is 10–15 °C higher, that of RNA:RNA hybrids around 20–25 °C higher, than for DNA:DNA hybrids given by the formula above. Base pair mismatches lower the melting point.

The kinetics of hybridization are decisively influenced by the diffusion rate and length of the duplexes; the diffusion rate is highest for small probe molecules. As a result, hybridization with oligonucleotide probes is usually complete in 30 min to 2 h, while hybridization with longer probes is typically carried out overnight. A disadvantage of oligonucleotide probes, however, is that their sensitivity is not as high as with longer probes, since both the length of the hybridizing sequence and the number of labels that can be incorporated are limited. Nonetheless, the sensitivity can be increased by the use of oligonucleotide cassettes (Section 28.2.1) and through terminal tailing with multiple labels.

Repetitive sequences hybridize faster since the number of potential matches for any given sequence is greater. Reaction accelerators, such as the inert polymers dextran sulfate or

poly(ethylene glycol), coordinate water molecules and thereby increase the effective concentration of the nucleic acids in the remaining solution.

The proportion of correctly paired nucleotides in a hybridized duplex molecule is determined by the degree of stringency with which the hybridization was conducted. Stringent conditions are those in which only perfectly base-paired nucleic acid double strands are formed and remain stable. When, under given conditions, oligo- or polynucleotide probes will only pair with the desired target nucleic acid in a mix of nucleic acids (i.e., no cross-hybridization with other nucleic acids takes place) the hybridization is defined as selective. An example of the use of this with oligonucleotide probes in specific hybridization is the detection of differences between almost identical sequences, with only a single base pair difference, such as *ras* wild-type/mutant at position 12 or the difference between *Neisseria gonorrhoeae* and *Neisseria meningitidis*, which only differ by a single base. Factors that influence the stringency are mainly ion concentration, the concentration of helix-destabilizing molecules, such as formamide, and the temperature. While monovalent cations, usually Na^+, and mutually repulsive, negatively-charged phosphates coat the helix backbone and thereby increase the stability of the double helices, formamide inhibits the formation of hydrogen bonds and thus weakens helix stability.

Temperature is of considerable importance. For example, the melting point, T_m, of a DNA segment composed of 50% (G + C), at an ion concentration of 0.4 M NaCl, is 87 °C. Hybridization takes place between 67 and 72 °C in this case. Adding 50% formamide lowers the melting point of the DNA helix to 54 °C, so hybridization can take place at 37–42 °C. This decrease in temperature is used in *in situ* hybridization, for example, because cellular structural integrity is lost at typical hybridization temperatures.

In Situ Hybridization, Section 35.1.4

Temperature is what defines the stringency of hybridization at defined formamide and Na^+ concentrations. The T_m of a duplex molecule decreases up to 5 °C for every 1% base pair mismatch; higher temperatures serve to allow only perfectly complementary sequences to pair (high stringency). By reducing the temperature, hybrids with unpaired bases are also tolerated (low stringency).

The higher the stringency, the more specific the hydrogen bonding between the complementary strands along the entire length of the hybridizing sequence. With oligonucleotide probes, it is possible to differentiate single mutations under stringent conditions, which is essential, for example, for the specific detection of point mutations, such as the single base pair difference in sickle cell anemia, or the detection of RNA sequences from certain pathogenic species of bacteria, such as *Neisseria gonorrhoeae*.

Use of high stringency conditions restricts the detection of hybridizing sequences to those that find a perfectly complementary match. After hybridization, the wash steps are conducted at only 5–15 °C below the T_m (destabilization of hybrid complexes) and in a buffer containing a low ion concentration (0.1 × sodium saline concentration (SSC), which corresponds to 15 mM Na^+).

A precise differentiation of mismatched base pairing is easiest using a PNA hybridization probe (Section 28.2.3). These artificial nucleic acid analogs, with an uncharged, peptide-like backbone, show more pronounced differences in stability than RNA or DNA probes between wild-type and mutant hybridizations, which results in better discrimination of base pair mismatches (also see LNA, Section 28.2.4).

28.1.3 Hybridization Methods

Heterogeneous detection systems employ a detection reaction subsequent to washing off the remaining probe. Homogeneous detection systems carry out detection without separating the remaining probe, which usually involves a change in probe state to turn a signal on when hybridized.

Heterogeneous Systems for Qualitative Analysis In addition to the Southern blots already mentioned, there are also dot, reverse dot, and slot blots methods for the qualitative analysis of DNA. The same methods can be used for RNA except that the Southern blot is called a Northern blot instead. Bacteria are detected with the aid of colony hybridization assays and viruses with plaque hybridization assays. Targets for *in situ* hybridization are chromosomes, cells, swabs, tissue, or even entire small organisms, such as *Drosophila,* on slides.

Heterogeneous Systems for Quantitative Analysis Heterogeneous systems for the quantitative analysis of nucleic acids include sandwich assays using capture and detection probes, replacement assays of a short detection probe from the complex, or special amplification methods, in which the labeling of the detection complex takes place by incorporation of a dNTP or a primer during amplification (e.g., with DIG; Section 28.4.3). The labeled amplicon is subsequently hybridized to a biotinylated capture probe and the complex immobilized on a streptavidin-coated membrane. Alternatively, reverse dot blots capture the target using a probe

covalently bound to the membrane. After washing away the excess label, the amount of bound, DIG-labeled amplicon reflects the original concentration of the target. In amplification assays, the principle is turned around and the primer is labeled with biotin, while detection is via hybridization with a labeled detection probe (Figure 28.4).

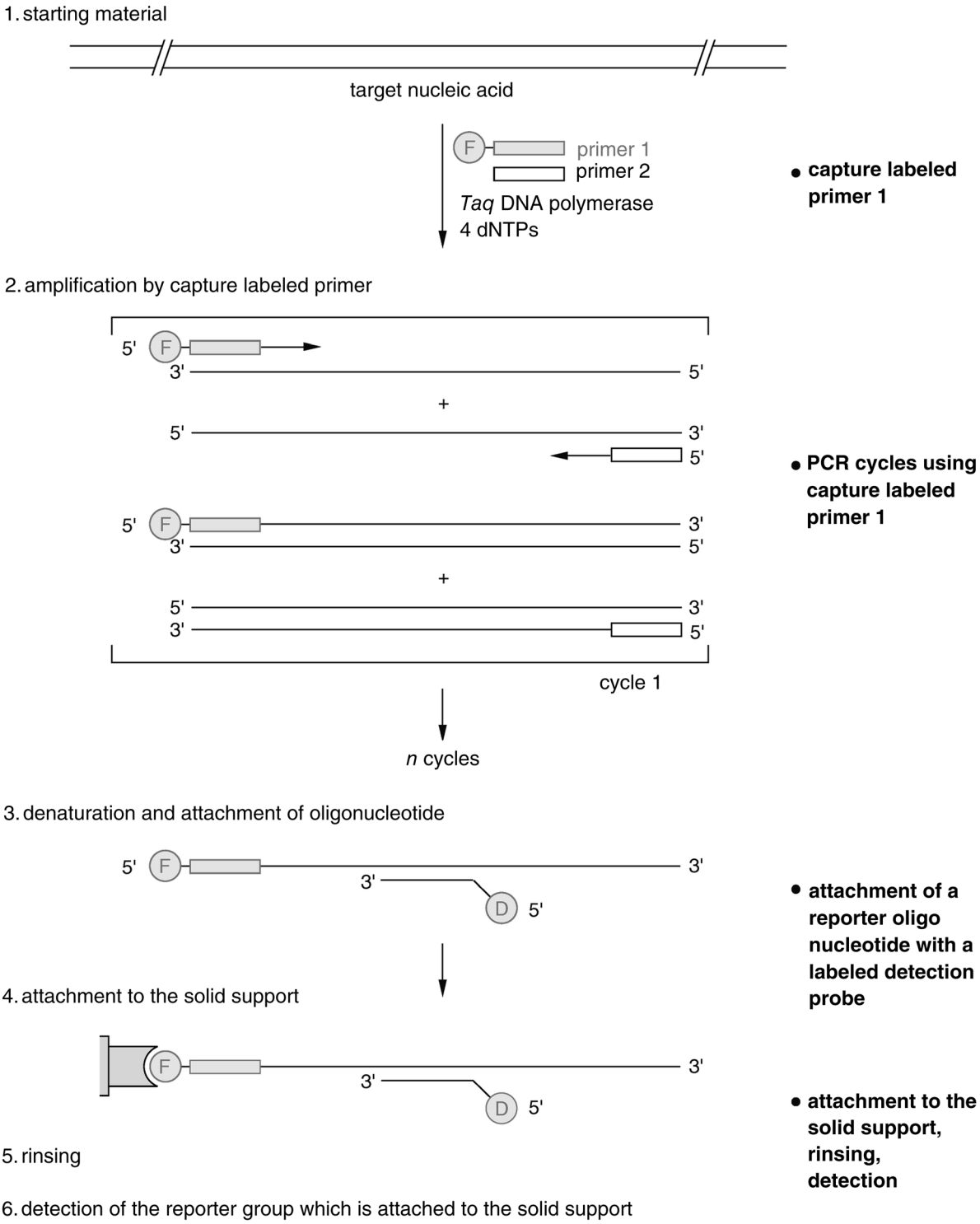


Figure 28.4 Principle of heterogeneous amplification systems. The capture marker attached to the solid support (F) is incorporated during amplification. The amplicon strand attached to the solid support is detected by hybridization with an oligonucleotide probe (D: detection marker).

Homogeneous Systems for Quantitative Analysis Homogeneous systems are much more difficult to design and develop but their convenience and large dynamic range gives them significant advantages over heterogeneous systems. In combination with efficient amplification techniques like PCR, homogeneous amplification techniques have enabled quantitative and reproducible detection of femto- to attogram amounts (10^{-15} to 10^{-18} g), which suffices to detect as little as a few copies of the target sequence. At this extreme level of sensitivity, statistical limitations in collecting the samples begin to limit the overall sensitivity of detection of the target sequences.

Amplification of DNA, Section 29.2.2

Homogeneous systems allow measurement of the amplification products during the course of the amplification reaction, without requiring the separation of reaction educts before addition of the detection reagent. The risk of contamination can be greatly reduced by carrying out the reaction in a closed system of sealed vessels, which also allows the direct detection of fluorescence signals at any time during the amplification reaction. The use of glass capillary tubes, for example, allows detection of fluorescent signals directly through the wall of the tube. The formation of the amplicons can be followed in real time. Another advantage of the homogeneous systems is their larger dynamic range, up to eight–to-ten orders of magnitude, in comparison to heterogeneous systems.

In homogenous systems, the probes bind to target sequences between the primers during the amplification, generating a detectable signal. This results either from digestion of the primer by the elongation enzyme, in the case of *TaqMan* probes, or through the binding itself, for *LightCycler* rapid PCR and 3′- and 5′ end-labeled probes (HybProbes), which hybridize next to one another. Molecular beacon-labeled probes open their branched structure on hybridization to the target sequence. This results from labeling the probes with a fluorescence-quencher set or two FRET pairs. Rhodamine or fluorescein are often used as the fluorescent marker in these systems and rhodamine derivatives, such as dabcyl or cyanine dyes, are used as the quencher. A particularly good quencher is the black hole quencher, which almost completely absorbs the emitted light. The type of label pair is tuned to the specific system.

Instruments, Section 29.2.1

***TaqMan* or 5′ Nuclease Methods** A modern homogeneous detection system is the *TaqMan* or 5′ nuclease amplification detection principle, also known as the 5′ nuclease assay. In this well-known homogeneous detection system the probe is equipped with a marker pair, consisting of a fluorescent marker and a quencher. The distance between the pair is chosen such that the incident primary light stimulates release of fluorescent light from the marker, which is absorbed by the quencher as long as the probe is free, preventing release of a detectable signal. After hybridization of the detection probe to the amplified target molecules, the 5′-3′ exonuclease activity of the *Taq* DNA polymerase frees the fluorescent nucleotide from the probe, which diffuses away from the quencher and can now emit unquenched light (Figure 28.5).

Prior to degradation by the 5′ nuclease catalysis, the probe's fluorescent signal is inactivated by the proximity of the quencher. As a result the free probe is inactive and need not be separated to specifically detect newly formed hybrid complexes. Since the quenched fluorescent nucleotide comes exclusively from hybridized complexes, the amount of fluorescence measured as a result of the decoupling of the fluorescent probe from the quencher is a direct measurement of the amount of the hybrid complex. This allows measurement of the intended target molecules without the need to separate the excess probes. Measurement of the signal increase allows the quantification of the amplified targets formed (Figure 28.6a).

Figure 28.7 shows the results of a *TaqMan* measurement: The amount of amplification product is not measured after completion of the amplification reaction, as in end point measurements, instead the formation of amplification products is measured continuously during the course of the PCR cycles, which is why it is called a real time measurement. The unit of measurement, the c_T^- value, gives the PCR cycle during which a signal is first seen above the threshold of detection. Plotting the c_T^- value semi-logarithmically against the initial copy number, prior to amplification, results in a linear relationship between the c_T^- value and the original copy number of the target sequence in the sample.

By correlating the results with corresponding curves from external controls, or co-amplification of the target sequence with internal standards of known copy number, the copy number of the target sequence in the sample can be quantitatively determined. Internal standards are constructed such that they can use the same primers but contain a different probe-binding sequence. If the probe for the desired target sequence is labeled with one dye, and the probe for the control with another, filters can be used to separate the two signals measured simultaneously.

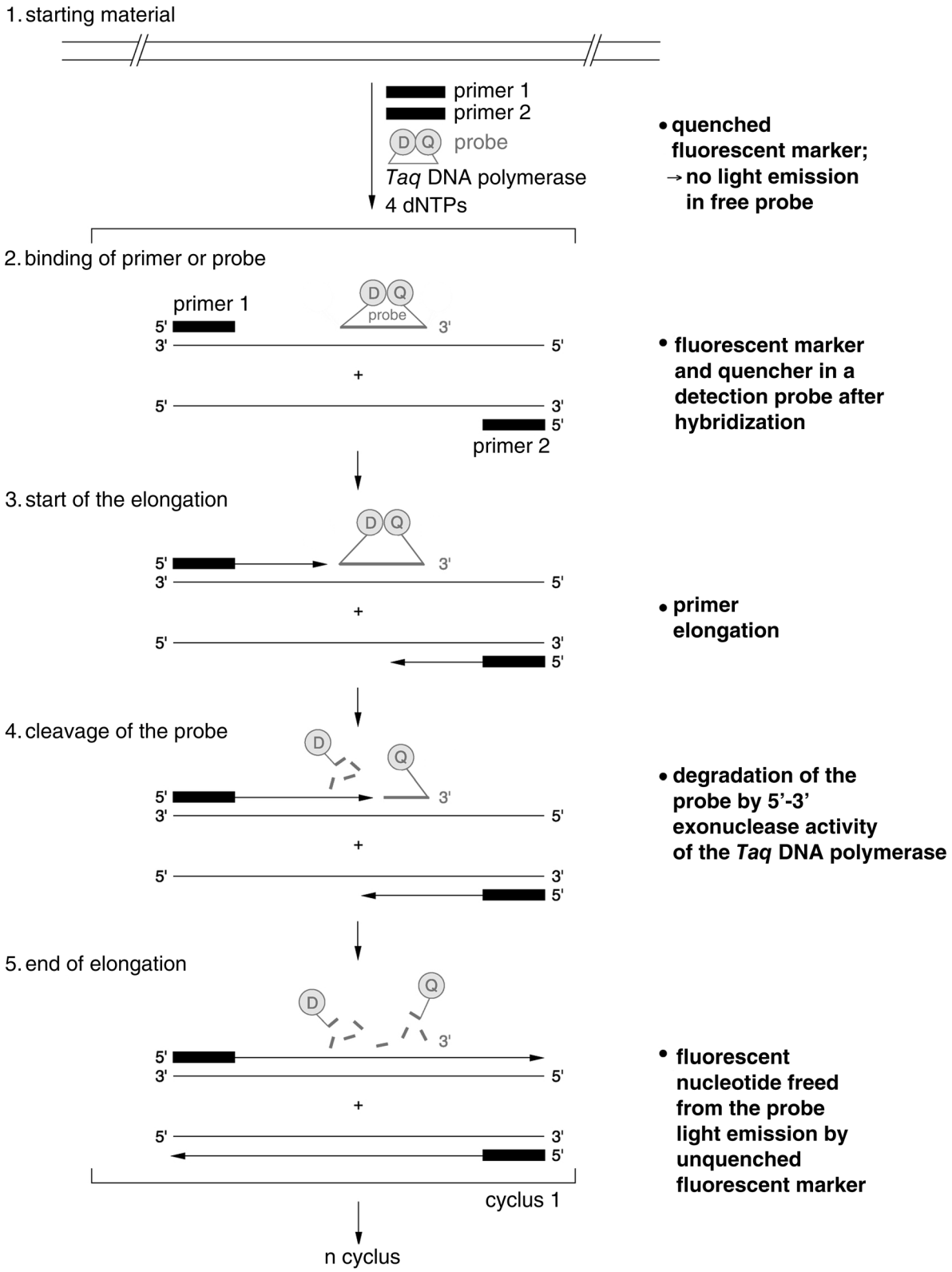


Figure 28.5 Principle of the 5′ nuclease reaction format (*TaqMan*). As long as the fluorescent detection marker (D) and quencher (Q) are linked in close proximity, no signal is emitted. The 5′ nuclease activity eliminates the linkage, the detector diffuses away and is no longer quenched, resulting in a signal proportional to the number of targets created by the *Taq* DNA polymerase.

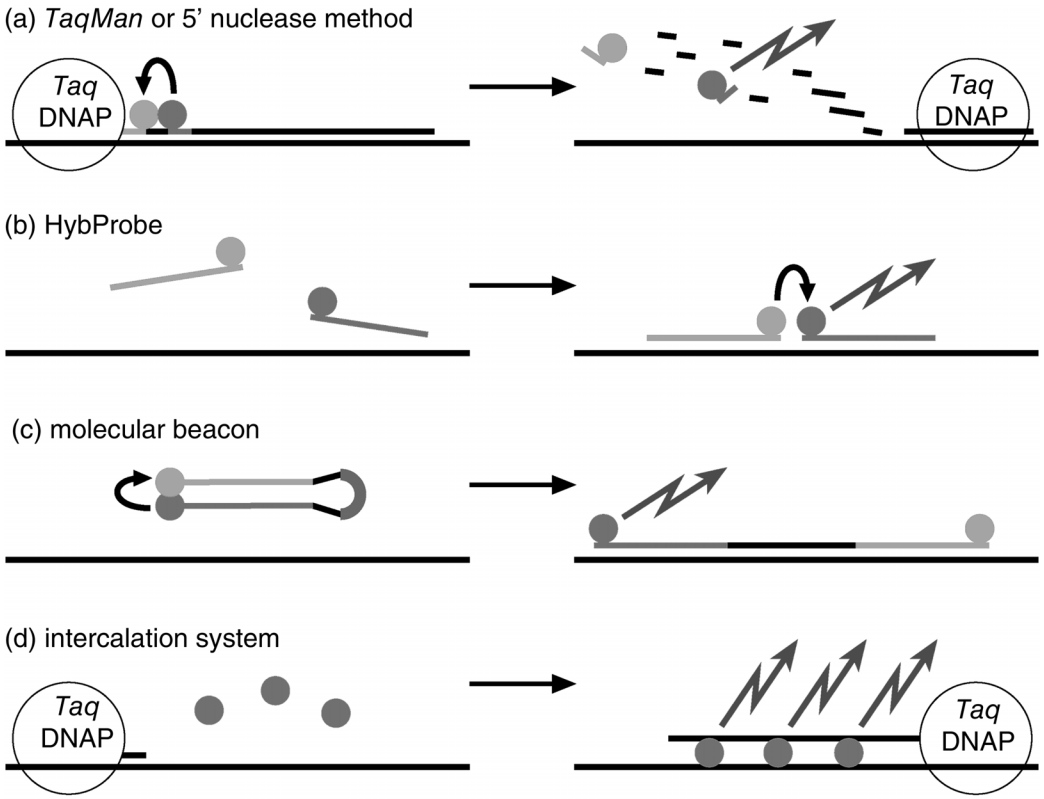


Figure 28.6 Homogeneous detection systems: (a) *TaqMan*/5′ nuclease system; (b) FRET system with HybProbes; (c) molecular beacon system; (d) intercalation system.

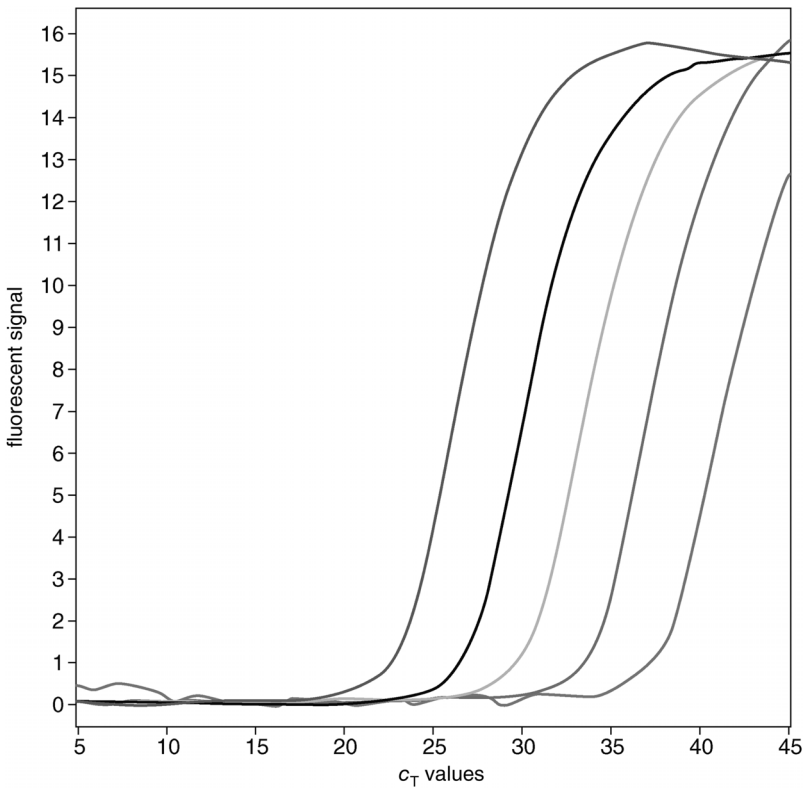


Figure 28.7 Coupled amplification and detection in the *TaqMan* system. The higher the c_T value, the lower the copy number of the sample shown by the curve.

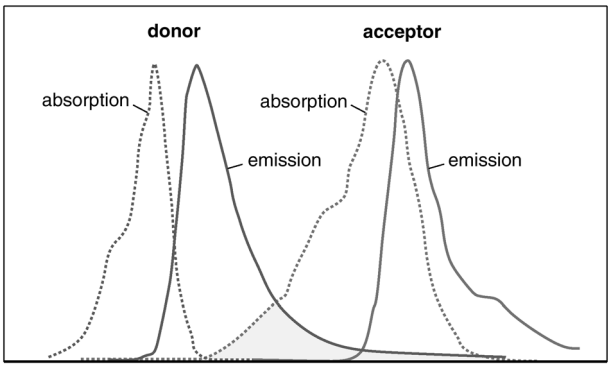


Figure 28.8 FRET principle. The absorption and emission spectra of the two FRET components are different. The emission spectrum of the energy donor, however, overlaps with the absorption spectrum of the energy acceptor, so that an energy exchange can take place. The detected secondary light emitted by the FRET pair is of a longer wavelength than the primary light input to induce the signal.

Förster Resonance Energy Transfer (FRET), Section 7.3.7

FRET System In the FRET system, probes are used that carry two fluorescent detection markers, with different but overlapping absorption and emission spectra. This allows energy to be transferred from one to the other, and also allows distinguishing between the emitted light signals. The two components form a fluorescence energy resonance transfer (FRET) pair. One component takes up the primary light and transfers it, in the form of light, to the second component, if it is close enough (0.5–10 nm). The second component emits the absorbed energy in the form of longer waved secondary light. The increase in wavelength allows the output signal to be measured without interference from the primary light (Figure 28.8).

HybProbes are probe pairs in which each of the two probe components are marked individually with one of the two FRET components. The upstream binding probe is marked on the 3′ terminal, the second probe, which binds immediately downstream, is labeled on the 5′ end. In solution, the two component probes do not result in a FRET signal since they are not in physical proximity. But if the two probes hybridize to the newly synthesized target sequence during the annealing phase of the PCR reaction they are now in direct proximity to one another. Since the FRET components are labeled on the directly proximal ends, the FRET effect between the components results in a signal. The more probe pairs that bind to the new amplicons, the stronger the signal becomes. Measuring the signal increase allows quantification of the amplified sequences (Figure 28.6b).

In the LightCycler, the FRET probes and samples are contained in thin glass capillary tubes, which allows rapid PCR temperature cycles. Probes used in the LightCycler are HybProbes, which have a fluorescent reporter and quencher pair. The tiny volume and high surface area of the capillaries allows for a very rapid temperature exchange. Each cycle is short, which reduces the total amplification time significantly.

Molecular Beacon System The use of molecular beacons is another way to measure the formation of products during amplification. In this case, in contrast to the HybProbes, only a single probe is used. The probe has a reporter–quencher pair bound to each end. An example of such a molecular beacon is the combination of fluorescein as the reporter and dabcyl as the quencher. The probe sequence is so chosen that the probe displays a pronounced stem-loop structure, which brings the photosensitive components on the two ends into immediate proximity of one another, leading to quenching. As a result, no fluorescent signal is observed. The other end of the stem-loop structure is covalently held together by the loop. The loop contains the sequence that can hybridize with the target sequence. When it does so, the probe is unfolded and the two ends are separated, which prevents the quencher from blocking the fluorescent signal of the detection marker.

Thus, molecular beacons are also a type of fluorescence dequenching assay. In this homogeneous system, the signal becomes stronger as more probe pairs bind to the formed amplicons, which directly reflects the amount of the amplified products. By measuring the increase in signal, this method allows quantification of the formation of amplified sequences (Figure 28.6c).

Intercalation Assay In dye intercalation assays the signal is generated by the intercalation of dyes like SYBR Green into the newly synthesized double strands of the amplification products. In

contrast to other assay types, no detection probe is hybridized to the amplicon, instead a sequence-independent, quantitative measurement is allowed by the intercalation of fluorescent dye into the amplification products (see also Section 28.3.1). Comparison of shifted LightCycler profiles of different products generated in a single run is possible if the amplicons differ in length or base composition and thus have different melting points.

Other Homogeneous Systems Other homogeneous systems are used primarily for the quantitative analysis of nucleic acids of bacteria or viral infections. Examples are the activation of inactive β-galactosidase by a complementary α peptide in enzyme complementation assays or the measurement of changes in mass of the complex through changes in fluorescence depolarization. These homogeneous assays will not be discussed in more detail.

***In Situ* Systems** *In situ* assays are used to analyze nucleic acids in fixed cells, tissues, chromosomes (metaphase chromosomes or prophase nuclei), or complete organisms such as whole mount embryos. Such molecular cellular analysis supplements the nucleic acid analysis to detect sequence or genetic aberrations in isolated nucleic acids discussed up to this point. *In situ* detection begins with the mounting of the biological material on slides or cover slips. Cell walls must be lysed enzymatically to allow hybridization with a labeled probe in the cells, whole mounts, or to the fixed chromosomes. The resulting optical (e.g., DIG-AP plus BCIP/NBT) or fluorescent signal (e.g., directly coupled fluorescein, rhodamine) creates the signal. An example of sequence detection in whole mounts is the detection of mRNA expression from early developmental genes in *Drosophila* embryos.

In Situ Hybridization, Section 35.1.4

In cells from higher organisms, DNA in the nucleus is in the form of chromatin or, during the metaphase of cell division, chromosomes. Cell division of many lymphoid cells can be stimulated with the application of phytohemagglutinin, a plant hormone, and cultivation for 2–3 days at 37 °C. The resulting metaphase chromosomes can be isolated by treating the cells with the spindle inhibitor colchicine; this freezes the cells in the middle of cell division. After pipetting the cells onto a glass slide, the cells are lysed with a hypotonic solution to create a chromosome spread. After fixing, the chromosomes can be observed under a microscope.

Fluorescence *in situ* hybridization (FISH) is the best-known system for karyotype analysis of chromosomes with the aid of fluorescence-tagged probes. Specific regions of the chromosomes can be detected with fluorescent signals.

Application of FISH and CGH, Section 35.2

The following sections go into the details of nucleic acid detection with a primary focus on hybridization assays and non-radioactive labeling and detection. Radioactive labeling is also discussed. Staining is presented in Section 28.4.1

28.2 Probes for Nucleic Acid Analysis

Labeled oligonucleotides or nucleic acid fragments play a central role in strategies for sequence-specific nucleic acid detection by both sequencing and hybridization by serving as primers for nucleic acid synthesis. Either radioactive or non-radioactive probes are employed. The repertoire of methods employed to generate and use non-radioactively labeled probes has expanded greatly in the last few years, making them the method of choice, particularly for standard methods of nucleic acid analysis. While radioactive methods were originally the only option, problems relating to contamination of sophisticated and expensive instrumentation and disposal problems have greatly favored the development of non-isotopic assays.

Although the routinely used isotopes ^{3}H, ^{14}C, ^{32}P, ^{33}P, ^{35}S, and ^{125}I have the advantage that the chemical structure, and therefore the hybridization characteristics, of the probes remains unchanged during nucleic acid analysis, the use of isotopes has the following serious disadvantages:

- limited half-life and therefore detection opportunity: for example, the frequently used ^{32}P isotope has a half-life of only 14.3 days;
- necessity of internal standards for quantitative analyses;
- the molecular damage to the probe itself caused by its radioactive emissions;

- need to repeat probe labeling for longer experiments;
- necessity for a special safety laboratory with expensive safety precautions;
- necessity of disposing the radioactive material;
- increased planning and logistics;
- potential health risks.

These disadvantages make the use of isotopes, particularly with the increasing availability of non-radioactive methods with at least comparable sensitivity and range of use, more problematic. Since, however, many laboratories in the research field still have the equipment for radioactive work, isotopes remain in use for blot hybridization and manual sequencing. It is, however, to be expected that with increasing standardization and automation of analytical methods the radioactive methods will be used less and less often.

Probe Types DNA and RNA probes, short single-stranded DNA oligonucleotides or longer, double-stranded DNA, or single-stranded RNA probes, are used in the analysis of nucleic acids. Cloned probes contain vector fragments, unless the vector sequences are removed in additional steps. Vector fragments can lead to undesired cross-hybridization; for example, unspecific cross-reactions of pBR vector sequences with genomic human DNA have been described. These undesirable effects can be avoided by using vector-free probes, which can be synthesized by PCR amplification, *in vitro* synthesis, or chemical synthesis.

A recently invented alternative to DNA oligonucleotides is peptide nucleic acid (PNA) oligomers, which contain the same base-specificity and hybrid geometry, attached to an uncharged peptide-like synthetic backbone. Due to the lack of the mutually repelling phosphate groups, hybrids of PNA probes and nucleic acid sequences have a higher melting point, allowing the use of higher hybridization temperatures and increasing the specificity of the hybridization. A further advantage of PNA is an increased ability to discriminate mismatches.

28.2.1 DNA Probes

The main types of DNA probes used in nucleic acid analysis are genomic probes, cloned DNA probes or corresponding restriction fragments, (i.e. DNA probes), PCR-generated amplicon probes, and synthetic oligonucleotides probes.

Polymerase Chain Reaction, Chapter 29

For many years, cloned cDNA or genomic fragments were the most commonly used DNA probes used to detect complementary DNA or RNA sequences in Southern or Northern blots. The probes were usually between 300 bp and 3 kb in length. The sensitivity depends on the length of the hybridizing region and the density of the labeling: as a result, genomic probes are more sensitive than cDNA probes, since the cDNA probes can only hybridize with the exon sequences of the nucleic acids to be detected, while genomic probes often encompass the extensive intron sequences. Both probe types, however, have the disadvantage that their cloning and subsequent plasmid isolation is time consuming. The production of vector-free probes requires an additional restriction cleavage and fragment separation. Even then, repetitive sequences within the section of the probe can lead to cross-reactions with eukaryotic DNA or amplified eukaryotic genes or cellular total mRNA. This can lead to unspecific bands.

The possibility of creating hybridization probes by PCR amplification with *Taq* DNA polymerase led to an enormous and rapid increase in the availability of probes. This procedure has several advantages:

- Cloning and plasmid isolation are no longer necessary, so the probes do not contain vector sequences.
- Both DNA and RNA can serve as templates to create probes (RNA must first converted into cDNA with reverse transcriptase in RT-PCR).
- The probes are of a defined length, allowing easy adjustment of the stringency of the hybridization.
- Probe design is extremely flexible, since the probe length and position can be easily controlled by the selection of the primers.
- Only the sequences where the primers bind need to be known for the amplification, thus it is possible to generate probes for new, unknown sequences between the primer binding sites.

- Similarly, probes for mutants with sequence variations in the probe area are readily available.
- The probes can be labeled during generation by using marked labeled nucleotides or primers, leading to a uniform density of labeling.
- By using new DNA polymerases or blends of polymerases, probes that are kilobases in length are possible, instead of the 100–1000 bp length from standard *Taq* DNA polymerases.

These advantages have made PCR the method of choice for the production of DNA probes. The probes can often be used directly in hybridizations. To avoid co-hybridization with unspecific amplification products, the amplification products are usually purified. Long PCR products can have secondary structure effects, resulting in lower sensitivity or unspecific hybridization signals; these can be avoided by additional restriction digestion.

Besides long PCR probes, synthetic oligonucleotides are increasingly used for hybridization probes. Modern oligonucleotide synthesis machines can create oligonucleotides up to a length of around 150 nucleotides with a defined sequence or with targeted sequence changes in any position. Oligonucleotide probes are well suited to detect point mutations. Oligonucleotides between 17 and 40 bp in length are used for this purpose, which allows the optimization of the hybridization and wash steps. Base pair mismatches are easiest to recognize when they are located in the middle of the hybridizing region; mutations in flanking sequences are not as well discriminated.

A further advantage of short oligonucleotide probes is that they hybridize faster than longer probes. A disadvantage is their lower sensitivity (Section 28.1.2). Nevertheless the strength of oligonucleotide probes does not lie in the detection of single copy genes nor low copy mRNA, but instead in the mutation analysis of PCR-amplified genes, strongly expressed mRNAs, or rRNA species amplified 10^3–10^4-fold.

DNA Microarray Technologies, Chapter 37

Oligonucleotide arrays are currently in development as hybridization tools for oligonucleotides and cDNA. These arrays contain a large number of oligonucleotides or sDNA capture probes with differing specificity. With these chips, a large number of mutations in different positions of the target amplicon can be analyzed in mutation/polymorphism analysis or from differing cells, tissues, or organs in expression pattern analysis.

28.2.2 RNA Probes

Single-stranded RNA probes are produced by run off *in vitro* transcription of sequences cloned into vectors containing bacteriophage SP6, T3, or T7 promoters (Figure 28.9). DNA fragments or PCR products are cloned into the multiple cloning site immediately downstream of the promoter. The recombinant vector is then cleaved directly at the 3′ end of the insert, which creates a fixed termination point for the run-off transcription. The strong promoter selectivity and the fixed

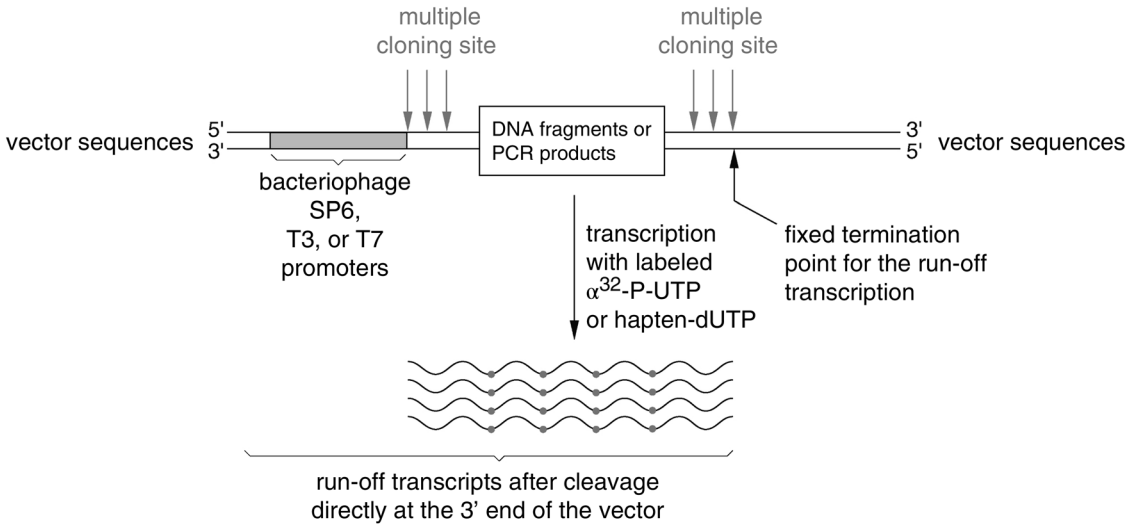


Figure 28.9 Synthesis of labeled RNA probes.

Template DNA and Detection of *in Vitro* Transcripts, Section 34.3.3

termination point lead to transcripts of identical length. The transcription cycle terminates and re-initiates 100–1000 times resulting in a high yield of the probe. Radioactive ribonucleotides or hapten-modified dUTP can be added during transcription to label the RNA probes in the process of transcription, in the same way as for PCR probes. New vectors contain two different promoters in opposite orientations on either side of the cloning region, which allow the transcription of complementary RNA strands with opposite polarities (sense/antisense RNA). To avoid unspecific hybridization signals from the vector segments, the *in vitro* transcripts are treated with RNase-free DNase. The main advantage of RNA probes is that DNA:RNA or RNA:RNA hybrid complexes have a higher melting point than DNA:DNA hybrid complexes. This provides increased sensitivity, so that low abundance mRNA in Northern blots or *in situ* can also be detected. However, RNA probes are vulnerable to ubiquitous RNases, so that all solutions and equipment must be sterilized with chemical additives like diethyl pyrocarbonate and heat treatment prior to use.

When used for *in situ* experiments, the run off transcripts are treated with a limited amount of RNase because full-length transcripts often penetrate the cell wall or membrane poorly; the shortened molecules penetrate better, which leads to more probe in the nucleus available for hybridization and, thus, increased sensitivity.

28.2.3 PNA Probes

Concept of Peptide Synthesis, Section 22.1

Synthetic peptide nucleic acids (PNAs) contain a peptide-like backbone which has many advantages relative to DNA and RNA probes (Figure 28.10). PNA probes can be produced in peptide synthesis machines as well as DNA synthesis machines. Boc synthesis chemistry is used for peptide analog synthesis; Fmoc synthesis chemistry is used for DNA analog synthesis. In both Boc and Fmoc chemistry protective groups are used – only the structural elements of the backbone differ. The solubility of PNA oligomers can be increased by the introduction of charged terminal or internal side chains (e.g., Glu, Lys) or charged groups in the labeling group spacers, so that the synthesis of up to 30-mers is possible. PNA oligomers have a range of advantages relative to DNA oligonucleotides:

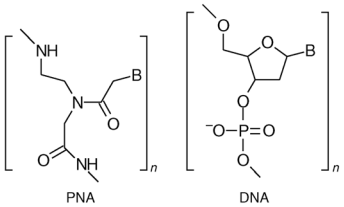


Figure 28.10 Structural comparison of PNA and DNA.

- Greater hybrid stability; therefore higher temperature, and correspondingly more stringent, hybridization conditions can be used.
- Shorter oligonucleotides have higher diffusion rates and faster reaction kinetics.
- Ion concentration-independent hybridization allows low salt concentrations: double-stranded PCR products hybridize without denaturing.
- The hybridization at low salt concentrations opens potential secondary structures within the target molecules.
- The T_m difference between matched and mismatched base pairs is more pronounced with PNA probes than with DNA or RNA probes. This allows better mismatch discrimination.
- The discrimination of mismatched bases is optimal throughout the entire length of the probe, except for the terminal three or four bases on the flanks.
- PNA probes are more resistant to nucleases and proteases due to the artificial structure of their backbones and base linkage, which increases the stability of the probes.
- The solubility of PNA can be increased by the use of charged amino acids (e.g., Lys, Glu) in the backbone or by charges in the marker linkers.

These advantages make PNA oligomers an attractive alternative to oligonucleotide probes for point mutation analysis. These advantages are particularly relevant to array systems on chips, since the selective detection of mismatches and avoiding secondary structure effects are of central importance in these systems. Also critical is the solubility of the PNA probes in these systems, which is achieved with a long linker molecule. PNA capture probes allow the selective isolation of target nucleic acids through the formation of very stable triplexes on certain target sequences or duplex formation in mixed target sequences.

28.2.4 LNA Probes

Locked nucleic acids (LNAs) are a new class of bicyclic DNA analogs in which the 2′ and 4′ position is coupled into a furanose ring by an O-methyl group (LNA: stabilization by bicyclic

bridged ribose; Figure 28.11). The binding of these analogs to complementary nucleic acids is the strongest of the known DNA analogs.

LNA probes are more hydrophilic, and therefore nuclease-resistant, than PNA, which makes them a desirable alternative for some purposes. In addition:

- LNA probes are more soluble than PNA probes, which makes their hybridization characteristics similar to DNA probes.
- The highest T_m values of any oligonucleotide analog pair are seen in LNA:DNA mixed oligonucleotides. This allows the use of shorter oligonucleotide probes for the determination of point mutations, which increases the discrimination between wild-type and mutation targets, similarly to PNA.
- The O-methyl bridge can also be replaced by a thio or amino bridge.

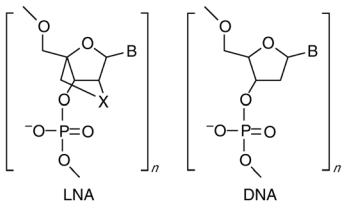


Figure 28.11 Structural comparison of LNA and DNA; X = O, S, NH.

28.3 Methods of Labeling

Non-radioactive modifications can be incorporated into probes in enzymatic, photochemical, or chemical reactions. Isotopes are usually incorporated into probes by enzymatic reactions. The labeling positions and type of label differs between methods, depending on whether isotopes or non-radioactive reporter groups are used.

DNA, RNA, and oligonucleotides can be labeled by means of the enzymatic incorporation of labeled nucleotides, resulting in high sensitivity probes that are densely labeled. The photolabeling of DNA and RNA results in less strongly labeled probes, but avoids damage caused by radioactive labeling, which can affect the length of the probe; thus, it is the most suitable method for the synthesis of labeled size standards. Chemical labeling was initially used to label DNA fragments and its use is increasing for the labeling of DNA or PNA oligomers. Figure 28.12 shows an overview of the most common non-radioactive labeling reactions, which are discussed in this section.

Enzymatic labeling uses either 5′ labeled primers (PCR labeling) or labeled nucleotides instead of, or in addition to, unlabeled nucleotides. Non-radioactive labeling uses nucleotide analogs modified with haptens like digoxigenin fluorescein or conjugates like biotin (Section 28.4.3). Hapten-dUTP or hapten-dCTP can be used as the enzyme substrate for RNA, DNA, and oligonucleotide labeling; the latter two can also make use of hapten-cATP. The labels can interfere with one another if they are too close to one another; consequently, maximum sensitivity requires a certain distance between them. The optimum distance needed to achieve highest labeling density and the necessary minimum separation is hapten-specific. The optimal labeling density is achieved with hapten-specific mixes of hapten-dNTP and non-modified dNTPs (e.g., 33% DIG-dUTP/67% dTTP).

High sensitivity requires that the modifications are not buried in the helix structure. Therefore, a spacer between the nucleic acid strand of the probe and the modifying group is critical. In the case of the haptens previously mentioned, the spacers are at least 11 atoms long and are often composed of oxycarbonyl elements, which are coupled via ester or amide bonds. The N and O atoms make the linker sufficiently hydrophilic.

28.3.1 Labeling Positions

Radioactive labeling involves exchanging stable natural isotopes for unstable radioactive isotopes, depending on the nature of the isotope, on different positions of nucleoside triphosphates. Exchanging isotopes does not change the chemical structure, so that the labeled molecules have the same chemical properties as the natural substances. This means that the reaction conditions do not have to be changed for enzymatically incorporating labeled nucleotides into probes, nor do the hybridization conditions require adaptation. The most commonly used labels are ^{32}P or ^{33}P phosphates, exchanged for either the α- or γ-phosphate residue in 2′-deoxyribo-, 3′-deoxyribo- (cordycepin) or 2′-ribonucleotides. Labeling with ^{35}S replaces an oxygen atom of the α-phosphate with the radioactive sulfur (Figure 28.13a). The ^{32}P- or ^{35}S-labeled α position remains attached to the nucleoside when probes are homogeneously labeled by polymerases (e.g., random-primed labeling, nick translation, reverse transcription, or PCR amplification), whereas end labeling, such as with T4 polynucleotide kinase, involves transfer of the labeled γ-phosphate from ATP to the free 5′ terminal OH of the

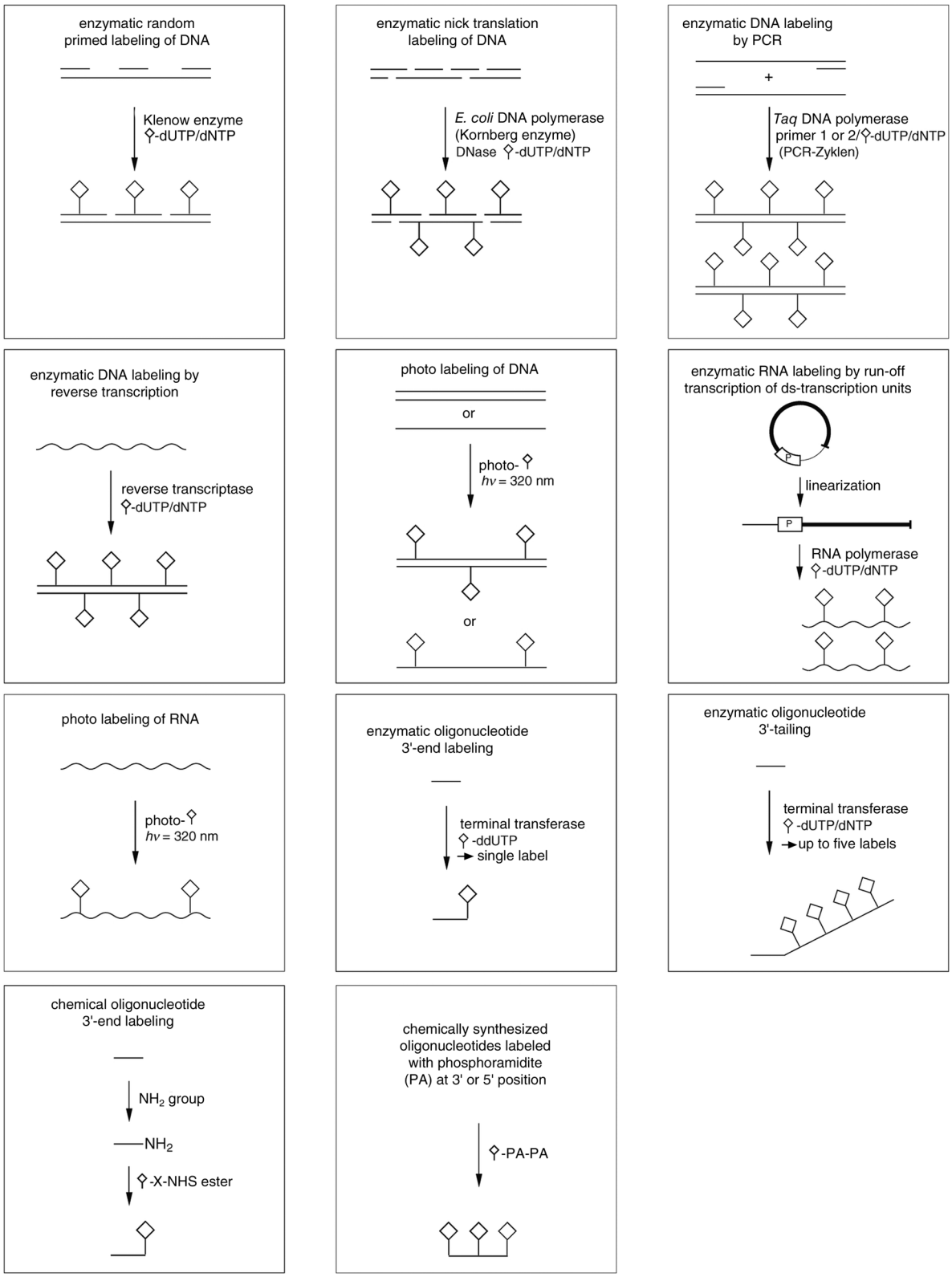


Figure 28.12 Schematic representation of the enzymatic, photochemical, and chemical labeling reactions, which are discussed in the following sections. Source: from Kessler, C. (1992) *Non-radioactive Labeling and Detection of Biomolecules*, Springer, Berlin, Heidelberg; and Kessler, C. (ed.) (2000) *Non-radioactive Analysis of Biomolecules*, Springer, Berlin, Heidelberg.

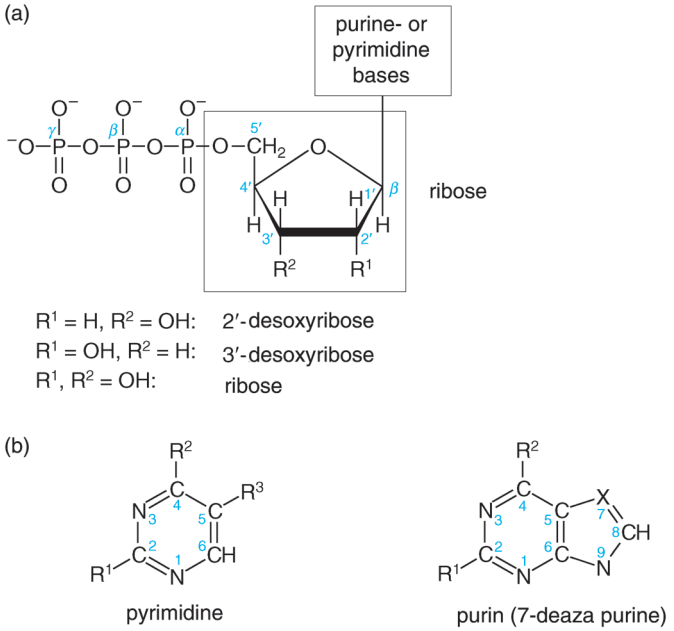


Figure 28.13 Exchange positions for radioactive labels: (a) nucleoside triphosphate; (b) positions of the base rings. For the residues R^1 to R^3, see the inside front cover; X = N, CH (7-deaza purine).

probe. The ^3H isotope is usually used for *in situ* applications, due to its lower radiation scatter and longer half-life. It is incorporated into various positions of the base ring. Labeling with ^{125}I is at the C5 position of cytosine (Figure 28.13b). The low radiation intensity and very long half-life, with its attendant disposal problems, of ^{14}C isotopes have made their use rare.

The chemical structure of the labeled nucleotides and the labeled probe is changed by the non-radioactive modification of probes with reporter groups. This means, for example, that the reaction conditions for enzymatic incorporation of the label need to be adapted to the altered substrate characteristics. Unspecific binding to the modifications must be blocked with suitable blocking reagents (e.g., milk proteins); in the case of blot hybridization, the entire membrane surface must be blocked; for *in situ* hybridization, the cell or tissue surface requires blocking. The correspondingly modified protocols are, however, well established and do not limit the use of non-radioactive probes.

The most commonly labeled position for sequencing primers, PCR primers, or hybridization oligomers (DNA, PNA) is the 5′-terminal. This preserves the characteristics of the probe and the formation of hydrogen bonds during hybridization is not influenced. Modifications are introduced through bifunctional linear diamino bonds of variable length. Nucleotides are usually labeled by base modifications; however, labeling of the 2′ position of ribose has also been described. The base modifications are chosen such that they do not interfere with the formation of hydrogen bonds during hybridization. The most common modification position is the C5 position of uracil or cytosine; in the case of deoxyuridine, the modification group imitates the methyl residue of the thymidine base of deoxythymidine (compare Figure 28.13b). In both cases the introduction of the modifying group is not direct, instead it is introduced attached to a spacer. Other suitable positions for the introduction of non-radioactive modifying groups are the C6 of cytosine and the C7 of deaza-guanine and deaza-adenine (Figure 28.13). The amino groups of cytosine, adenine, or guanine, which were frequent targets in the past, are less well suited, since these positions are involved in hydrogen bonding.

28.3.2 Enzymatic Labeling

Many of the enzymatic labeling reactions are analogous when using radioactive or non-radioactive labels. The difference is, however, that in the case of radioactive labeling enzyme substrates are structurally identical isotope-labeled nucleotides, while non-radioactive labeling involves additional modification of the nucleotides, which may require adaptation of the reaction conditions.

Enzymatic Labeling of DNA Homogeneous DNA labeling involves random priming with the large fragment of the *Escherichia coli* DNA polymerase I (Klenow enzyme), nick translation with *Escherichia coli* DNA polymerase I (Kornberg enzyme) or via PCR amplification with *Taq* DNA polymerase. The density of the labeling is around one label per 25–36 base pairs. Oligonucleotides can be labeled with the help of the terminal transferase reaction; depending on the substrate, one to five labels per oligonucleotide are attached (for an overview see Figure 28.12).

Random Priming In the random priming protocol, double-stranded DNA is denatured and rapidly cooled and high concentrations of primer are added to prevent the rehybridization of the target strands. The primers are a mix of all 4096 conceivable hexanucleotides (thus "random" primer), so that statistically every target sequence is covered and the hybridization can occur at any point in the sequence. The Klenow enzyme, the large subtilisin fragment of the *Escherichia coli* DNA polymerase holoenzyme, extends the primer in a template-dependent reaction. During the elongation reaction, unlabeled dNTPs and hapten-modified dUTPs are built in. Since the template strand has been replicated, strand displacement leads to a new round of synthesis, which leads to a yield of over 100% of the input template DNA. Since statistically several primers bind per target, each primer elongation only replicates part of the sequence; the result is a mix of probes of variable probe length. The partial sequences are all target-specific, however, and carry a homogeneous labeling. Random primed DIG-labeled probes reach a high detection sensitivity in the sub-picogram range.

Nick Translation Nick translation uses the *Escherichia coli* DNA polymerase holoenzyme (Kornberg enzyme) and very small amounts of pancreatic DNase I. This technique requires the 5′-3′ exonuclease activity of the DNA polymerase, as well as its polymerase activity. The DNase catalyzes the creation of single strand nicks; a very precisely controlled, small amount of DNase is required to ensure that the nicks remain limited. Both ends flanking the single-strand break function as templates, either for the 3′-5′ polymerase activity or the 5′-3′ exonuclease activity. The 5′-3′ exonuclease activity successively removes 5′-phosphorylated nucleotides, while the 3′-5′ polymerase activity simultaneously fills in the holes with new, labeled nucleotides. By this concerted action, the nick moves in the 5′-3′ direction, referred to as nick translation. It involves a DNA replacement synthesis, the yield though remains below 100%, which is less than random priming. Besides the lower labeling, getting the relationship between the amounts of DNase I and *Escherichia coli* DNA polymerase exactly right is a critical factor. As a result, nick translation is going out of fashion for labeling probes.

Polymerase Chain Reaction, Chapter 29

PCR Amplification As previously mentioned, PCR amplification with *Taq* DNA polymerase is a useful method for the creation of vector-free probes. The amplification reaction consists of 30–40 temperature cycles with three partial reactions – denaturation, primer binding, and primer elongation – each at different temperatures. Besides this three-step protocol, two-step protocols have also been described, in which primer binding and primer elongation take place in a single step. Labeled primers or labeled dNTPs are used in the creation of labeled products as hybridization probes. PCR amplification is explained in detail in the next chapter.

Reverse Transcription Labeled DNA probes can also be synthesized from RNA targets with viral reverse transcriptases (e.g., AMV RTase, MMLV RTase). After first and second strand cDNA synthesis, unlabeled dNTPs and hapten-labeled dNTPs are added to the reaction mix to label the probe.

Terminal Transferase Reaction DNA oligonucleotides can be labeled enzymatically by the template-independent attachment of labeled dNTPs, sometimes referred to as tailing, with the enzyme terminal transferase. If a mix of labeled and unlabeled dNTPs are employed, tails with multiple labels are created in a template-independent reaction. Using labeled cordycepin triphosphate (3′-dATP) or 2′,3′-ddNTPs allows the attachment of only a single labeled nucleotide, since the reduced 3′ position can no longer be extended. (see Figure 28.12: enzymatic oligonucleotide 3'-end labeling, oligonucleotide 3'-tailing).

Template DNA and Detection of *in Vitro* Transcripts, Section 34.3.3

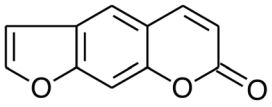
Enzymatic Labeling of RNA RNA can also be labeled using the previously described *in vitro* run off transcription with bacteriophage-encoded SP6, T3, or T7 RNA polymerases (Section 28.2.2). Owing to the re-initiation of transcription, high synthetic yields are achieved (up to

20 µg transcript from 1 µg recombinant vector). The density of the labeling for enzymatic DNA labeling is around one label for every 25–36 nucleotides.

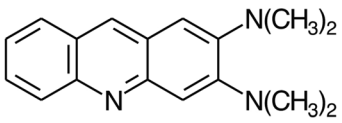
28.3.3 Photochemical Labeling Reactions

There is only one photochemical DNA and RNA labeling reaction: aryl azide-activated haptens are reacted with nucleic acids under long wavelength UV light. Light excitation leads to the release of elementary nitrogen (N_2); the short-lived reactive nitrogen radical reacts with different positions of the DNA or RNA and thus covalently bonds the hapten to the nucleic acid. The labeling density is relatively low, however, at around one label per 200–400 base pairs.

A range of photoactive substances intercalates into nucleic acids and can be subsequently covalently bound to nucleic acid bases in a photoreaction. These substances include coumarin compounds (psoralen, angelicin), acridine stains (acridine orange), phenanthroline (ethidium bromide), phenazine, phenothiazine, and chinone (Figure 28.14). For photolabeling the most important of these is the bifunctional psoralen, which forms either mono- or bis-adducts with pyrimidine bases after intercalation and photoactivation.



psoralen



acridine orange

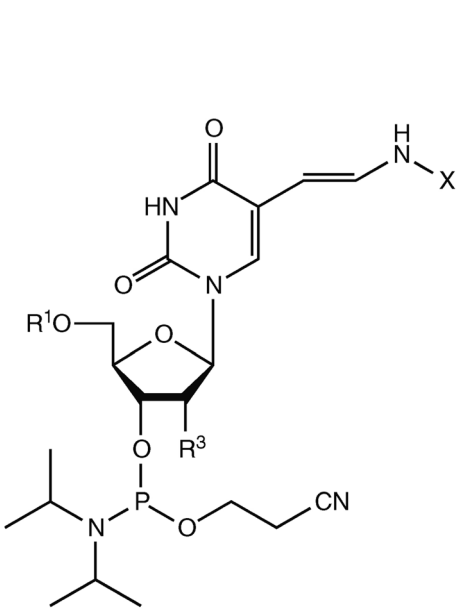
Figure 28.14 Intercalating, photoactive substances for the detection of nucleic acids. The compound first intercalates and a subsequent photoreaction bonds it to the nucleic acid strand covalently.

28.3.4 Chemical Labeling

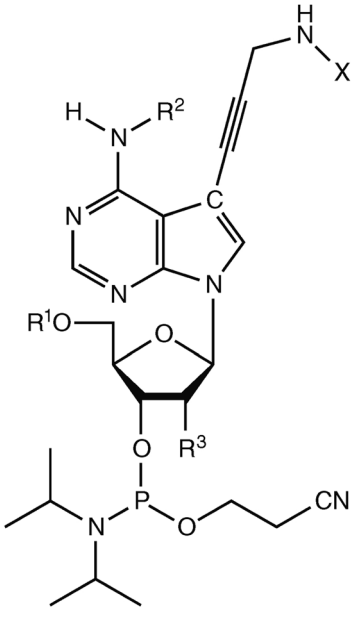
Chemical labeling reactions are used mainly for the labeling of DNA, LNA-oligonucleotides, and PNA-oligomers. Labeling takes place during solid-state oligonucleotide synthesis by the direct incorporation of modified phosphoramidites. Alternatively, labeling can take place after synthesis by coupling the modifying group and the 5′ terminal phosphate of the oligonucleotide with the help of bifunctional diamino compound linkers. Protected phosphoramidites containing the desired hapten are commercially available. Such labeled oligonucleotides can often be used directly after removal of the protective group, without need for further HPLC purification.

Principles of the Synthesis of Oligonucleotides, Section 27.6.1

A second possibility is the incorporation of uracil or cytosine phosphoramidites carrying protected allyl- or propargylamine residues onto the C5 position or deaza-purines onto the C7 position during nucleic acid synthesis (Figure 28.15). After removal of the protective group, they react with *N*-hydroxysuccinimide-activated haptens or other coupling-capable reporter



allylaminouracil phosphoramidite



propargylamino-7-deazaadenine phosphoramidite

Figure 28.15 Examples of modified phosphoramidites used in oligonucleotide synthesis. Shown are a pyrimidine and a purine nucleoside as DNA and RNA building unit. R^1 and R^2 are protecting groups, R^3 = H, OH, and X = hapten, reporter molecule.

molecules. In this case, subsequent purification by gel electrophoresis or HPLC is often necessary, owing to the often non-quantitative synthetic yield.

Oligonucleotides can be coupled directly to labeling enzymes like alkaline phosphatase or horseradish peroxidase with bifunctional linking reagents. Direct detection of the enzyme is possible after hybridization with such enzyme-tagged probes; however, the hybridization temperature and duration are limited by loss of enzyme stability.

After synthesis and removal of the protective groups, the chemical labeling of PNA involves coupling of the 5′ terminal amino group with the hapten with the aid of the previously described bifunctional diamino compounds.

28.4 Detection Systems

Radioactively and non-radioactively labeled probes, primers, or nucleotides are the central components of nucleic acid detection systems for sequencing and hybridization. The hybridization formats allow greater flexibility in the use of labels and detection components than sequencing protocols. This is because direct and indirect systems, in which a particular modifying group can be detected and measured in multiple ways, can be used. This allows great variety in the application of non-radioactive detection systems, not only in blot formats but also in the full range of qualitative and quantitative systems.

If the detection of particular sequences is unnecessary and the pattern of restriction fragment patterns is of interest, staining with intercalating compounds like ethidium bromide is the method of choice. The formation of nucleic acids can also be detected immediately in scintillation counters after incorporation of radioisotopes and separation of unincorporated labeled nucleotides; while these methods were once in widespread use, they are now only rarely used.

28.4.1 Staining Methods

Determination of Nucleic Acid Concentration, Section 26.1.4

The absorption of UV light can be used to determine the concentration of nucleic acids; double-stranded DNA can also be detected by the intercalation of fluorescent dyes or by silver staining. Although silver staining is more sensitive, the visualization of DNA fragments in gels is routinely done by intercalation with ethidium bromide and illumination with UV light to generate a fluorescent signal. The sensitivity of the ethidium bromide dye depends on the length of the fragments, like all staining methods, and lies in the nanogram range. The sensitivity of silver staining is in the sub-nanogram range.

28.4.2 Radioactive Systems

The β-emitters ^{32}P, ^{33}P, and ^{35}S are usually used for blotting procedures with radioactive probes. Although ^{35}S isotopes have approximately tenfold less energy, they have the advantage of a longer half-life. ^{33}P isotopes are more expensive, but offer more energy than ^{35}S and better resolution than ^{32}P, making them a good compromise in some situations. ^{3}H isotopes are also β-emitters but have tenfold less energy than ^{35}S. Owing to its high stability and its low scattering, this isotope is used *in situ* and in tissues, though these techniques are increasingly switching to non-radioactive methods. Scattering refers to the tendency of high energy emitters to expose film, even when striking the film at an oblique angle, which causes the signals to become so-called diffuse reflections that appear blurry. Trichloroacetic acid (TCA) precipitations have been used to employ ^{14}C, but today this isotope is only of historical significance. Table 28.1 lists the key facts for probe labeling and sequencing.

The highest energy isotope in common use is ^{32}P, which makes it the most sensitive. However, its short half-life and high scattering limits the use of this isotope in sequencing and the analysis of complex fragment patterns, since the resolution of closely packed bands is limited. The legibility of sequence gels is also limited towards the origin. When high resolution is more important than high sensitivity, ^{33}P or ^{35}S are better choices. Examples are enzymatic sequencing and the analysis of DNA- or RNA-binding proteins in gel shift assays, due to altered

Table 28.1 Characteristics of radioisotopes used to label probes.

Isotope particle		E_{max} (MeV)	Half-life	Application	Characteristics
^{3}H	β	0.0118	12.3 years	*In situ*	Low sensitivity, high resolution
^{35}S	β	0.167	87.4 days	Filter hybridization, sequencing, *in situ*	Medium sensitivity, good resolution
^{125}I	γ	0.035	60.0 days	*In situ*	Low sensitivity, high resolution
^{32}P	β	1.71	14.2 days	Filter hybridization, sequencing	Highest energy, highest sensitivity, medium resolution due to reflection

mobility of the protein-bound fragments in gels. Besides ^{3}H, the low radiation intensity of ^{35}S and ^{125}I make them suitable for *in situ* applications, where limiting β-scatter is important to determine exact cellular localizations.

Radioisotopes are commercially available in the form of nucleotides already labeled at the desired position. The aqueous solutions contain stabilizers to inhibit the degradation of the biologically active substances by the ionizing radiation. The isotopes are stored at –20 or –70 °C to balance thermal heat decay. Owing to the steady degradation and formation of radicals, radioactive nucleotides should be used as soon as possible, even when the half-life allows a longer use.

The most important factors with respect to the storage of radioactively labeled probes are:

- the half-life of the isotope;
- the specific radioactivity of the probe; probes with a high labeling density are very sensitive but are subject to rapid degradation;
- the position of the radioactive atom in the molecule; internal labeling leads to strand breaks more easily than terminal labeling.

The detection of radioactive nucleic acids takes place in blot form by autoradiography with X-ray film, which can be stored permanently to document the results. There are different means of detection, depending on the isotope and the required sensitivity:

- *Direct autoradiography:* The radiating surface (membrane, gel, cell layer or tissue slice) is brought into direct contact with the X-ray film. This method applies to all β-emitters. Film without a protective layer is required for ^{3}H, so that the energetically weak electrons can penetrate to the photoactive layer.
- *Fluorography:* the radiating surface is treated with fluorescing chemicals, which convert the radiation energy into fluorescence; the most common fluorophores are 2,5-diphenyloxazole (PPO) and sodium salicylate.
- *Indirect autoradiography with intensifier screens*: High energy β-emissions are absorbed by phosphate residues of the intensifier screen and converted into visible light by illumination with a laser.
- *Fluid emulsions for cytological or cytogenetic* in situ *applications:* The low to mid energy ^{3}H or ^{35}S decay products require direct contact with the detection medium; the solid emulsion is melted at 45 °C and the slide is dipped into it. After drying, exposure occurs for days to months at 4 °C in the absence of light.
- *Pre-exposed X-ray film for direct autoradiography and fluorography:* A short pre-exposure activates the silver grains, which then require fewer photons to generate a signal. The pre-exposure can only be used for fluorography or light intensifiers (light processes).

The right method of detection depends on the characteristics of the isotope in use, such as its type, specific radioactivity, and total radioactivity, as well as the requirements of the experiment, such as the required sharpness, and maximum feasible exposure time.

28.4.3 Non-radioactive Systems

The non-radioactive labeling and detection systems are divided into direct and indirect indicator systems (Figure 28.16).

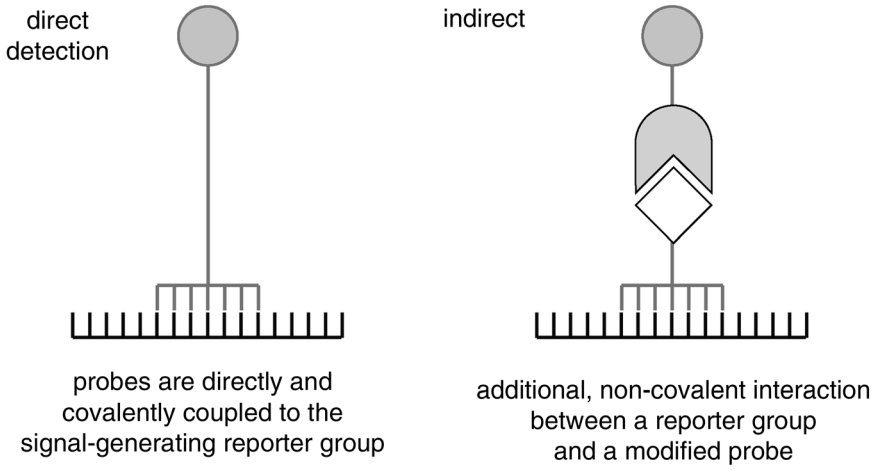


Figure 28.16 Direct and indirect detection systems.

The two types of reactions differ in the number of components and the reaction steps, as well as the flexibility of use. While direct systems are primarily used in standardized processes (e.g., for the labeling of universal sequencing primers), the more flexible indirect systems are used to selectively detect nucleic acid sequences.

In direct systems, the probes are directly and covalently coupled to the signal-generating reporter group; detection is done in two reaction steps:

- hybridization between the nucleic acid target and the directly labeled probe;
- signal generation by the directly coupled reporter group.

The advantage of direct systems is that only hybridization between the target and the probe is needed. The disadvantage is, however, that for every hybridization sequence each probe must be covalently coupled to the detection label. Therefore, this detection method is used primarily for the labeling of standard sequences with easily coupled fluorescent dyes. , e.g. with sequencing primers.

In indirect systems, the probes are not directly labeled, instead they are detected by an additional, non-covalent interaction between a low molecular weight tag and a universal detector. Thus, indirect systems first require the enzymatic, photochemical or chemical incorporation of the modifying group into the probe (Section 28.3). These incorporation reactions are easy; the corresponding protocols are well-established. A universal detection unit binds to the tag, independent of the type of probe and its specificity. The detector, which contains a binding unit as well as the reporter, couples specifically and with high affinity to the tag of the probe. The indirect detection takes place in three reaction steps:

- hybridization between the nucleic acid target and the modified probe;
- specific and high affinity, non-covalent interaction between the modifying groups of the probe and binding components of the universal detection unit;
- signal generation by the indirectly bound reporter component.

Although an additional reaction step is required, the high flexibility in the generation of the probes and the coupling with diverse types of detectors are significant advantages for indirect systems. As a result, simple and fast reactions can be used to build different tags into different types of probe; in addition, the tags can be detected with a large set of alternative, universal detectors, dependent on the application. The additional non-covalent interaction makes many combinations possible that allow a broad use of the non-radioactive reporter systems in basic research and the applications previously described at the beginning of the chapter.

Direct Detection Systems The most commonly used non-radioactive reporter groups in direct detection systems are fluorescent or luminescent reporter groups, as well as reporter enzymes. Gold-labeling is used for *in situ* applications; additional use of colored latex beads or silver staining amplifies the signal up to 10^4-fold. Box 28.1 shows an overview of important reporter groups.

Box 28.1 Important direct, non-radioactive reporter groups.

Fluorescent label

Direct fluorescence

- Fluorescein
- Rhodamine
- Hydroxy-coumarin
- Benzofuran
- Texas Red
- Bimane
- Ethidium/TB^{3+}
- Time-resolved fluorescence
- Lanthanide (Eu^{3+}/Tb^{3+}) micelles/chelates

Fluorescence energy transfer (FRET)

- Fluorescein (FAM)
- Rhodamine (TAMRA)
- Cy 3
- Cy 3,5
- CY 5
- CY 5,5
- CY 7

Luminescence labels

Chemiluminescence

- (Iso-)luminol derivatives
- Acridinium esters

Electroluminescence

- Ru^{2+} (2,2′-bipyridine)$_3$ complexes

Luminescent energy transfer

Metal label

Gold- labeled antibodies

Enzyme labels

Direct enzyme coupling

- Alkaline phosphatase
- Horseradish peroxidase
- Microperoxidase
- β-Galactosidase
- Urease
- Glucose oxidase
- Glucose-6-phosphate dehydrogenase
- Hexokinase
- Bacterial luciferase
- Firefly luciferase

Enzyme channeling

- Glucose oxidase: horseradish peroxidase

Enzyme complementation

- Inactive β-galactosidase: α-peptide

Polymer labels

- Latex particles
- Polyethyleneimine

Marker enzymes are such as alkaline phosphatase (Jablonski, E. *et al.* (1986) *Nucl. Acids Res.*, **14**, 6115–6128), horseradish peroxidase (Renz, M. and Kurz, C. (1984) *Nucl. Acids Res.*, **12**, 3435–3444; as well as fluorescent tags such as fluorescein or rhodamine (Kessler, C. (1994) *J. Biotechnol.* **35**, 165–189), and Kessler, C. (ed.) (2000) *Non-radioactive Analysis of Biomolecules*, Springer, Berlin, Heidelberg.

Bacterial alkaline phosphatase (AP) is mainly used for the direct labeling of oligonucleotides and horseradish peroxidase (POD) for the direct labeling of fragments. The use of marker enzymes requires an additional substrate reaction (see below).

Coupling alkaline phosphatase to oligonucleotides is a single-step reaction using a bifunctional linker. Direct AP-coupled oligonucleotides are useful in standard assays with a fixed sequence. For example, AP-coupled primers are employed in sequencing, direct blotting electrophoresis (DBE), and as universal amplifiers in signal amplification systems, like probe brushes (Section 28.5.3). The use of POD labeled fragment probes is limited, since POD is increasingly unstable above 42 °C, which limits the maximum temperature to a range that is not suitable for all purposes.

Well-known fluorescent labels are fluorescein, rhodamine, and coumarin derivatives. Higher sensitivity is achieved with phycoerythrins or fluorescein lattices; in these cases the coupling

FISH Analysis of Genomic DNA, Section 35.2.1

reactions are more complex. Besides use as sequencing primers, fluorescent markers are mainly used for fluorescent *in situ* hybridization (FISH).

The detected fluorescence can contain unspecific signals due to unspecific background light or fluorescent reaction components. For example, hemoglobin fluoresces, which causes interference in any experiments carried out using serum. This can be avoided with time-resolved fluorescence measurements with europium or terbium complexes, chelates, or micelles coupled directly to the probe via a linker, since the emission of the secondary light is delayed in these cases.

Direct luminescence markers can be grouped according to their activation type, which can be chemical, electrochemical, or biochemical. Well-known chemically activatable markers for the direct measurement of nucleic acids are acridinium esters, which are activated by H_2O_2/alkali, as well as the protein aequorin from the jellyfish *Aequorea*, which is activated by Ca^{2+} ions. Acridinium esters glow, as they release photons over a long time frame. Aequorin flashes a short, intense pulse of light that is very specific due to the extremely low background, leading to high sensitivity.

Electrochemical luminescence markers are stimulated to emit photons by electrochemical reactions. Corresponding markers are $[Ru^{2+}(bipyridyl)_3]$ or phenanthroline complexes. The ruthenium ions are oxidized on a gold electrode ($Ru^{2+} \rightarrow Ru^{3+}$) while the subsequent reduction of the Ru^{3+} by tripropylamine (TPA) creates a chemiluminescent signal. The resulting Ru^{2+} ion is then available to begin the next reaction cycle.

Gold particles can be used in blot or *in situ* formats to directly visualize targets. An additional silver staining can increase the sensitivity of the detection. In this case, the original gold particles have silver layered on them, increasing their size and making them easier to see.

Indirect Detection Systems Several different couplings are available to indirect detection systems due to the additional specific interaction between the modified nucleic acid hybrid and the universal detector carrying the reporter group. Table 28.2 shows the coupling groups commonly used for indirect nucleic acid detection. Most systems use antibodies or the biotin-binding proteins avidin or streptavidin to recognize special modifying groups attached to nucleic acids, but other less widespread systems exist that use specific sequences in the hybrid or specific hybrid conformations as binding components (operators or promoters) to bind to proteins like repressors or RNA polymerases. The DIANA concept rests on the binding of *lac* repressor β-galactosidase conjugates to hybrid-coupled *lac* operator tag sequences.

In addition, non-sequence specific binding proteins like the single-stranded binding (SSB) protein or histones have been employed as tags. Conformation-specific antibodies are examples of conformation-recognizing binders. Tagging probes with metal ions or to poly(A)-coupled systems were also used in the early days of the development of non-radioactive reporter systems.

Of the various systems, only the biotin (BIO) system and antibody systems with digoxigenin (DIG), fluorescein (FLUOS) and 2,4-dinitrophenol (DNP) have sensitivity in the sub-picogram range. Thus, these systems have become standards for the non-radioactive detection of nucleic acids, while the other systems are described more out of historical interest. Enzymatic labeling takes place with hapten-labeled nucleotides; as examples of these labels, the structures of DIG-, FLUGS- and BIO-labeled nucleotides are shown in Figure 28.17a. The labeling of

Table 28.2 Interaction pairs for indirect, non-radioactive detection systems. Source: from Kessler C. (ed.) (1992) *Non-radioactive Labeling and Detection of Biomolecules*, Springer, Berlin, Heidelberg; for further inetraction pairs see Kessler, C. (ed.) (2000) *Non-radioactive Analysis of Biomolecules*, Springer, Berlin, Heidelberg.

DNA modification ↔ binding partner	Examples
Vitamin ↔ binding protein	Biotin ↔ streptavidin
Hapten ↔ antibodies	Digoxigenin ↔ anti-digoxigenin antibody
Protein A ↔ constant region of IgG	Protein A ↔ IgG
DNA/RNA-hybrid ↔ DNA/RNA specific antibody	DNA/RNA ↔ anti-DNA/RNA antibody
RNA/RNA-hybrid ↔ RNA/RNA specific antibody	RNA/RNA-hybrid ↔ anti-RNA/RNA antibody
Binding protein-DNA-sequences ↔ binding protein	T7-promotor ↔ *Escherichia coli* RNA polymerase
Heavy metal ↔ sulfhydryl-reagent	Hg^{2+} ↔ HS-TNP ↔ (TNP/DNA)-specific antibody
Polyadenylation-polynucleotide phosphorylase/ pyruvate kinase	ATP-coupled red firefly luciferase reaction

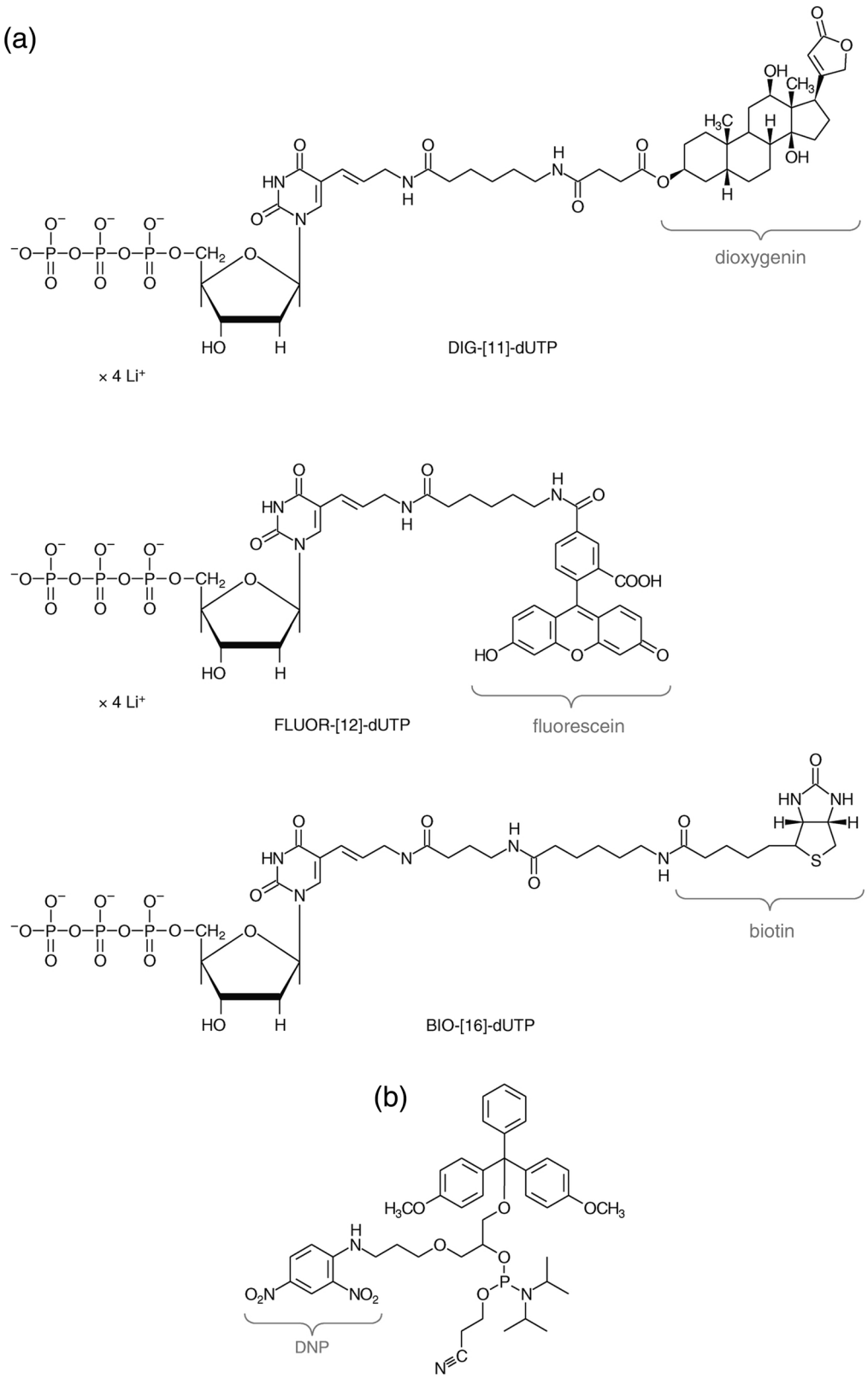


Figure 28.17 (a) Structure of hapten and biotin-labeled dNTPs. (b) DNP-modified phosphoramidites.

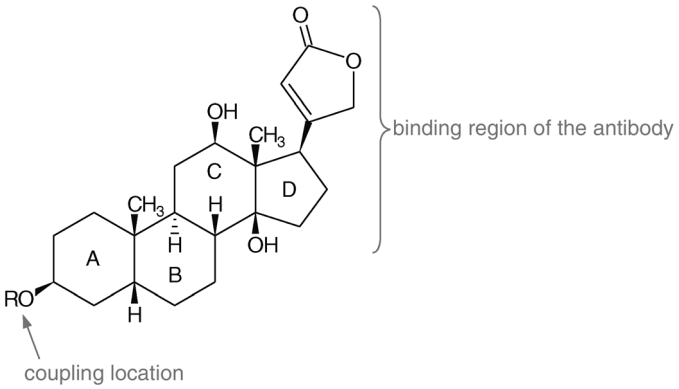


Figure 28.18 Structure of digoxigenin. Digoxigenin is a steroid with the formula $C_{23}H_{35}O_5$. The A/B rings are cis isomers, the B/C rings are trans isomers, and the C/D rings are cis isomers.

oligonucleotides takes place mainly through modified phosphoramidite; an example, the DNP-modified phosphoramidite, is shown in Figure 28.17b.

FISH Analysis of Genomic DNA, Section 35.2.1

Mixtures of differently labeled probes are used for the parallel detection of different fragments on blots (DIG, BIO, FLUGS: rainbow detection) or *in situ* for the detection of different chromosomal segments or different chromosomes (DIG, BIO, DNP: multiplex FISH, chromosome painting).

The Digoxigenin System Digoxigenin is the chemically synthesized aglycone of the cardenolide lanatoside C (Figure 28.18). The digoxigenin:anti-digoxigenin (DIG) system is based on the specific interaction between digoxigenin and a high affinity, DIG-specific antibody covalently bound to a reporter group. The DIG system can specifically detect sub-picogram amounts

Table 28.3 Important optical, luminescent, and fluorescent detection types. Source: from Kessler, C. (ed.) (1992) *Non-radioactive Labeling and Detection of Biomolecules*, Springer, Berlin, Heidelberg; for further detection types see Kessler, C. (ed.) (2000) *Non-radioactive Analysis of Biomolecules*, Springer, Berlin, Heidelberg.

Format	Specific detection		
	Optical	Luminescence	Fluorescence
Blots	AP/BCIP, NBT	AP/AMPPD, Lumiphos™, CSPD®, CDPstar®	Fluorescein, rhodamine, hydroxycoumarin, AP/AMPPD, fluorescein/rhodamine/hydroxycoumarin
	AP/naphthol-AS-azo-dye, diazonium salt	β-Gal/AMPGD	β-Ga/AMPGD, fluorescein/rhodamine/hydroxycoumarin
	HRP/TMB, immunogold		
Solution	AP/*p*-NPP	AP/AMPPD, Lumiphos™, CSPD®, CDPstar®	AP/4-MUF-P
	β-Gal/CPRG POD/ABTS™	β-Gal/AMPGD, POD/luminol, isoluminol	β-Gal/4-MUF-β-Gal, POD/homovanillic acid-*o*-dianisidine/H_2O_2
	GOD:POD-pair/ABTS™	Xanthine-oxidase/cyclic dihydrazide	Aromatic peroxalate compounds/H_2O_2
	Hexokinase: G-6-PHD-pair ABTS™	G-6-PDH/phenazinium salt	Lanthanoid-($Eu^{3+}Tb^{3+}$)-complex,-micelles-, chelates
		Hydrolysis of acridinium ester, rhodamine:luminol pair, AP/D-luciferin-*O*-phosphate: firefly-luciferase/ATP/O_2 *Renilla*-luciferase green fluorescent protein, $Ru^{2+}(bpy)_3$	
In situ	AP/BCIP, NBT		Fluorescein
	POD/TMB		Rhodamine
	Immunogold		Hydroxycoumarin

ABTS: 2,2-azino-di(3-ethyl)-benzothiazole sulfate
AMPGD: 3-(4-methoxyspiro(1,2-dioxetan-3,2′-tricyclo[3.3.1.1^{3,7}]-decan)-4-yl)phenyl-β-galactoside
AMPPD: 3-(4-methoxyspiro(1,2-dioxetan-3,2′-tricyclo[3.3.1.1^{3,7}]-decan)-4-yl)phenyl-phosphate, Na_2
BCIP: 5-bromo-4-chloro-indolylphosphate
CPRG: chlorophenol-red β-galactoside
CSPD: 3-(4-methoxyspiro(1,2 dioxetan-3,2′,5′-chloro)-tricyclo[3.3.1.1^{3,7}]-decan)-4yl)-phenylphosphate Na_2
Naphthol-AS: 2-hydroxy-3-naphtholicacid-anilide
Diazoniums salt: Fast blue B, Fast Red Tr, Fast Brown RR
GOD: glucose oxidase
MUF: methylumbelliferyl phosphate
NBT: nitro-blue tetrazolium salt
NPP: *p*-nitrophenyl phosphate
POD: horseradish peroxidase
TMB: 3,3′,5,5′-tetramethyl-benzidine

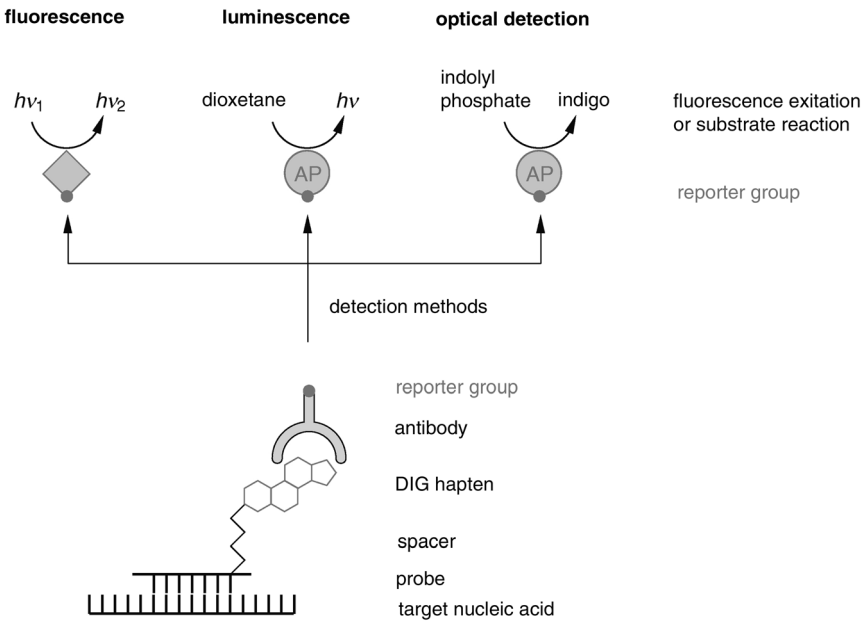


Figure 28.19 Digoxigenin detection system with its detection marker alternatives: fluorescent, luminescent, and optical.

of DNA or RNA. Alkaline phosphatase is often used as the reporter group, which is covalently conjugated to the antibody. Alkaline phosphatase catalyzes the conversion of optical or chemiluminescent substrates (BCIP/NBT, AMPPD; Table 28.3 below). The DIG system is shown in Figure 28.19. Enzyme-linked immunosorbent assays (ELISAs) employ detection with universal antibody-marker enzyme conjugates.

Immune Binding, Section 5.3.3

The high specificity of detection and the low background level of the DIG system are due to the fact that digoxigenin is naturally only present in *Digitalis* plants (foxglove) in the form of lanatoside compounds; the antibodies employed thus do not recognize any cellular component from other biological materials. This is particularly important for *in situ* hybridizations. DIG-specific antibodies have very few unspecific cross-reactions with cellular components. Only human serum contains components that are known to interact with digoxigenin-specific antibodies; these serum components can be specifically removed, however, by pretreatment of the polyclonal anti-DIG antibodies with serum.

Due to the high specificity of the digoxigenin antibodies used, even the structurally similar steroids are only recognized to a very limited extent or not at all; examples are the bufadienolide K-strophanthin (cross-reactivity $< 0.1\%$) and the steroids estrogen, androgen, or progestogens (cross-reactivity $< 0.002\%$). The important difference between digoxigenin and the sexual hormones is that the C and D rings are in the cis isomer conformation, rather than trans. Digoxigenin is coupled to the nucleotide via the –OH group to the C3 position of the cardenolide frame with a linear spacer. This does not interfere with antibody binding so detection with antibody conjugates is still possible after incorporation of digoxigenin into the probe.

Digoxigenin is isolated from leaves of the plant *Digitalis lanata* or *Digitalis purpurea* by the cleavage of three digitoxose and a glucose unit of the natural substance deacetyl-lanatoside C.

To further reduce unspecific effects, only the Fab fragment of the antibody is used, rather than the complete antibody. The Fab fragment is isolated by papain cleavage of the constant Fc fragment; it contains just the short antibody arms with the highly variable binding sites. The complete antibody is only used in coupled systems with secondary reporter antibodies, which recognize the Fc portion. For example, a secondary antibody from mouse, which recognizes the Fc portion of the DIG-specific sheep antibody, serves to amplify the signal. The reporter groups are not coupled to the primary DIG-specific antibody, but instead to the Fc-specific secondary antibody.

The Biotin System With the biotin:avidin (or streptavidin) (BIO) system, the ubiquitous biotin, also known as vitamin H, is used as the tag (for the structure of biotin see Figure 28.17a). Coupling

employs linear spacers attached to the terminal carboxy group of the biotin side chain. Avidin, from egg white, or streptavidin, from *Streptomyces avidinii* bacteria, have four high affinity binding sites for biotin; their binding constants are among the highest known natural affinities, at 10^{15} mol^{-1}. Avidin or streptavidin is covalently coupled with a reporter group to allow detection. After binding to the biotin-modified hybrid complex, secondary biotinylated detection components can be coupled to the remaining free biotin binding sites to amplify the signal.

Although the BIO system is as sensitive as DIG, the disadvantage of the system is that biotin is ubiquitously present in biological material. As a result, unspecific binding of free endogenous biotin is possible with almost all cellular material. This results in a high background, particularly for *in situ* formats. Another problem with this system is the tendency of the two binding proteins to stick to membranes, even after the proteins have been saturated with blocking reagents. This stickiness is caused by two factors: the high basicity of avidin and the presence of multiple tryptophan residues in the binding pockets of avidin and streptavidin. Both factors lead to unspecific polar or hydrophobic interactions with proteins of the membrane blocking reagents. In the case of avidin, unspecific binding is also caused by glycan chains on the surface, which can bind to sugar-binding proteins on the surface of cells (for *in situ* procedures) or the membrane blocking reagents (in blotting procedures). Since streptavidin is isolated from bacteria, this factor plays no role, since there is no protein glycosylation in prokaryotes.

These unspecific background reactions can be reduced in several ways:

- Acetylation or succinylation of the lysine chains or complex formation between avidin and the acidic protein lysozyme reduces the basicity of avidin and thus its unspecific polar interactions.
- Deglycosylation of avidin reduces unspecific adsorption by binding to sugars.
- Pre-incubation of the blocked membrane in a high ion concentration buffer reduces unspecific protein interactions.

Despite the high binding affinity and the other measures to reduce the unspecific background, a poor signal-to-background ratio is observed at low target concentrations, which limits the sensitivity of the entire system.

While this may matter when using BIO in a detection system, for other applications this is less important. The biotin–streptavidin binding system is often employed in the isolation of nucleic acids by hybridization with biotin capture probes. Samples are coupled to a streptavidin-coated surface (e.g., beads or microtiter plates) and the unbound analytes are washed away. The sequence-specifically bound nucleic acids can subsequently be amplified and detected, for example, after incorporation of DIG, specifically and without interference.

The Fluorescein System The fluorescein:anti-fluorescein (FLUOS) System (Figure 28.17a) is another antibody-based system. The hapten, fluorescein, is coupled by amide binding between the spacer and the free carboxy group of the hapten. The sensitivity of the fluorescein-specific antibody is also high. Since fluorescein is light-sensitive, however, exposure to light during storage causes loss of sensitivity. This can be avoided by storage in a cooled place protected from light.

The Dinitrophenol System The 2,4-dinitrophenol:anti-2,4-dinitrophenol (DNP) system is also based on antibody binding (Figure 28.17b). The hapten is bound to the aromatic C1 position, for example, by chemical conversion of 1-fluoro-2,4-dinitrobenzene with a protected amino terminal spacer into DNP-labeled phosphoramidites. These are then built into oligonucleotides during chemical synthesis. Since DNP is a synthetic compound it also has few unspecific interactions with biological material. DNP-labeled oligonucleotides are often used for more complex *in situ* investigations, like multiplex FISH.

Non-radioactive Detection in Indirect Systems In indirect detection systems, the entire repertoire of detection systems can be conjugated to the component that binds the tag on the hybridizing nucleic acid, depending on the format. Besides the reporter groups used in direct detection systems (fluorophores, luminescent dyes, gold atoms, see above), marker enzymes are often coupled to the tag-binding component, which create an optically visible luminescent or fluorescing reaction product through a catalytic substrate reaction. Well-known marker enzyme systems are bacterial alkaline phosphatase (AP), horseradish peroxidase (POD), β-galactosidase (β-Gal), luciferase, and urease.

The alternatives for detection can be summarized as follows:

- *optical systems:* mixtures of indolyl derivatives and tetrazolium salts or diazonium salts for coupled redox reactions;
- *chemiluminescent systems:* dioxetane derivatives, (iso-)luminol derivatives, acridinium esters, aequorin;
- *electrochemiluminescent systems:* Ru^{2+} or phenanthroline complexes;
- *bioluminescent systems:* luciferase derivatives;
- *fluorescent systems:* Attophos, Eu^{3+} or La^{3+}-complexes, -micelles, or -chelates;
- *FRET systems:* fluorescein and rhodamine derivatives;
- *fluorescence quencher systems:* fluorescein (FAM) and rhodamine (TAMRA) derivatives as the fluorescing component, Black Hole Quencher (BHQ), TAMRA-, dabcyl-, cyanine-derivatives as quenchers.
- *metal precipitating systems:* silver deposition on antibody-bound gold atoms (immunogold);
- *electrochemical systems:* urease-catalyzed pH changes;
- *in situ detection systems:* fluorescence *in situ* hybridization (FISH), primed *in situ* hybridization (PRINS), chromosome painting, multiplex FISH (M-FISH with SKY probes), comparative genome hybridization (CGH).

In addition, the following amplification reactions can be coupled to the primary signal generator (Section 28.5):

- *probe crosslinking:* generation of crosslinked structures (e.g., Christmas trees, probe brushes);
- *crosslinking of binding components:* examples include polystreptavidin, polyhapten, PAP, APAAP;
- *poly-enzymes:* crosslinked marker enzymes, for example, poly-AP;
- *signal cascades:* for example, $NAD^+/NADH + H^+$ cycles coupled to a redox color reaction that recycles the substrate for the next round.

Table 28.3 summarizes the most important systems for optical, luminescent, and fluorescent detection.

Optical Detection Optical enzymatic detection systems are based on the conversion of colored substrates, coupled with a change in the wavelength of absorption. Either colored precipitates (blot or *in situ* applications) or colored solutions (for quantitative measurements) are used. The most important substrates for colored precipitates are mixtures of 5-bromo-4-chloro-indoxyl phosphate (BCIP, Figure 28.20) or β-galactosidase (X-gal) and nitro-blue tetrazolium salt (NBT, Table 28.3), giving rise to a deep blue–violet precipitate (indigo/diazonium salt) after cleavage of the phosphate in a coupled redox reaction.

Luminescence Detection Luminescent systems are based on the chemical, biochemical, or electrochemical activation of substrates, which emit light as the atoms return to their ground state.

Chemiluminescence Common substrates like AMPPD or AMPPG (Table 28.3) are used for the detection of chemiluminescence; they form intermediate dioxetanes that decompose under chemiluminescence. After enzymatic cleavage of the phosphate or β-galactoside residue, the dioxetane coupling becomes labile.

The AMPD anion is created in an unstable, excited state and it decays by emitting light at a wavelength of 477 nm (Figure 28.21).

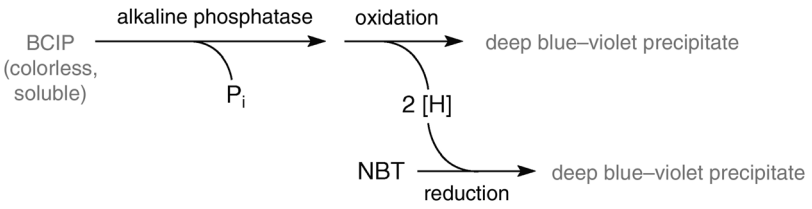


Figure 28.20 The coupled optical redox reaction BCIP/NBT.

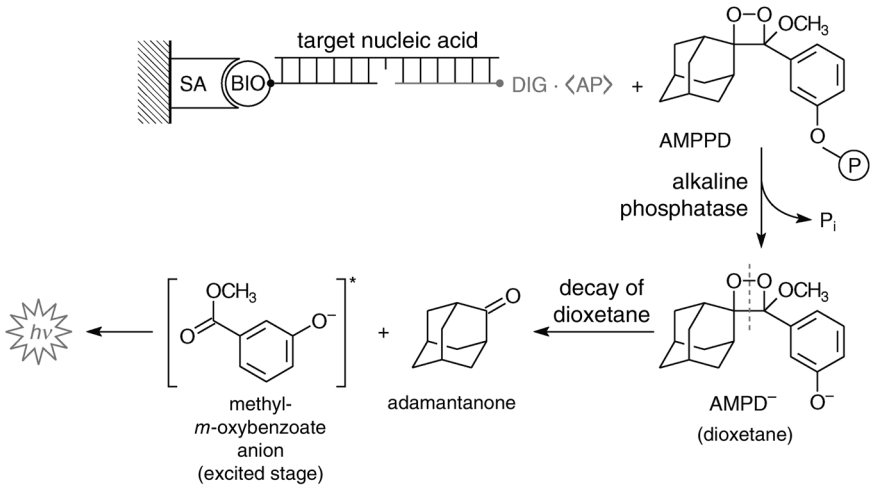


Figure 28.21 Mechanism of the dioxetane chemiluminescence reaction. The target nucleic acid is anchored to a solid support by a biotin–streptavidin (SA) interaction, thus marking it with the DIG/AP system. AP then activates the chemiluminescent substrate.

Several derivatives differ in terms of the stabilizing moiety and the stabilizers in the substrate solutions (e.g., Limiphos™, CSPD®, CDPstar®). The different formulations lead to different rates of decay or to different light intensities. The light emission can also be modulated by the addition of fluorescent detergents, which surround the dioxetane molecules like micelles. Depending on the type of additive, blue, red, or green secondary fluorescence (aquamarine, ruby, emerald) is created.

POD catalyzes the oxidation of luminol. The sensitivity of the chemiluminescence is increased in enhanced chemiluminescence by the addition of certain phenols (e.g., *p*-iodophenol), naphthols (e.g., 1-bromo-2-chloronaphthol), or amines (e.g., *p*-anisidine).

Electrochemiluminescence In addition to their use in direct detection systems, Ru^{2+} complexes (Section 28.4.3) are also used in indirect systems, after coupling with hapten-specific antibodies, for chemiluminescent detection after electrochemical excitation.

Bioluminescence Owing to their high sensitivity, luciferase enzymes from fireflies (*Photinus pyralis*) or bacteria are used. The eukaryotic enzyme catalyzes the conversion of luciferin into oxyluciferin. In an AP-coupled indicator reaction, D-luciferin-*O*-phosphate is the substrate for AP. After cleavage of the phosphate, luciferase converts the formed D-luciferin in the presence of ATP, O_2, and Mg^{2+} ions into oxyluciferin, thereby emitting light (Figure 28.22).

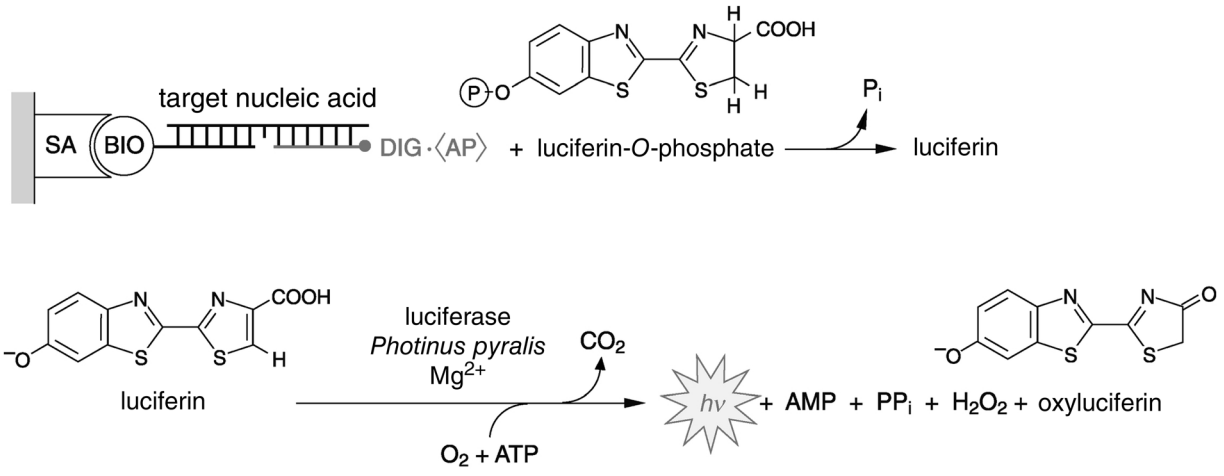


Figure 28.22 Mechanism of the luciferin bioluminescence reaction. Analogous to the chemiluminescence reaction in Figure 28.21, the target nucleic acid is anchored to a solid support by a biotin: streptavidin (SA) interaction marking it with the DIG:AP system. AP then activates the bioluminescent substrate.

Alternatively, a combination of the enzyme glucose-6-phosphate-dehydrogenase, NAD(P)H-FMN-oxidoreductase, and bacterial luciferase can be used. The $FMNH_2$ formed in the redox reaction is oxidized in the presence of decanal and O_2, resulting in the emission of light.

Green Fluorescent Protein (GFP), Section 7.3.4 and 34.4.3

Renilla luciferase, from sea anemones, catalyzes the bioluminescent oxidation of coelenterazine. In the presence of green fluorescent protein (GFP) this leads to green secondary fluorescence at a wavelength of 508 nm. The *Renilla* enzyme is often used indirectly as a secondary reporter molecule in the form of a biotin conjugate.

Fluorescence Detection Fluorescent light is created by absorbing primary light (broadband or monochromatic laser light); the excited fluorescent molecule returns to its ground state by emitting longer wavelength secondary light. Because of the possible overlap of the incident primary light and emitted secondary light, background signals can occur. Background effects are reduced by the geometry of the detector (perpendicular detector orientation) or by time-resolved fluorescence (TRF) with Eu^{3+}- or Tb^{3+}-complexes, micelles, or chelates (Section 28.3.3). Coupling several of these TRF fluorophores to the four binding sites via indirect coupling by biotin:streptavidin amplifies the resulting signal.

Indirect fluorescence can be detected by coupling the previously described direct fluorescence markers (e.g., fluorescein or rhodamine) to hapten binding antibodies (e.g., anti-DIG-fluorescein or anti-DIG-rhodamine conjugates).

FRET Detection Probes marked with interacting fluorescent components are used for FRET detection. The two components form a fluorescence energy resonance transfer (FRET) pair. One component absorbs the primary light and transfers it in the form of energy to the second component, if it is in immediate physical proximity (Figure 28.8). The second component emits the absorbed energy in the form of longer wavelength secondary light. The longer wavelength light is selectively measured to separate the signal from the input light used to excite the FRET pair.

Förster Resonance Energy Transfer (FRET), Section 7.3.7

A modern homogeneous detection system is the 5′-nuclease system (*TaqMan*). In this well-known homogeneous detection system the probe is labeled with a marker pair consisting of a fluorescent marker and a quencher (Figure 28.5).

For HybProbes two directly adjacent probes are used, marked with a FRET pair at their vicinal ends.

In the case of molecular beacons one probe carrying a FRET pair on both ends (termini) is sufficient; see Figure 28.23 for FRET components: fluorophores, and quencher (c, D).

***In Situ* Detection** Fluorescein and rhodamine derivatives are usually employed as the fluorescent markers for *in situ* detection. These are directly or indirectly bound to the probes. DIG, biotin, or dinitrophenol (DNP) are used as the interacting components for indirect binding. For SKY probes, direct or indirect probe labeling with different marker combinations (Orange, Texas Red, Cy5, Spectrum Green, and Cy5.5 or DIG:Cy5.5 or FITC, Cy3, Bio-Cy3.5, Cy5, DIG-Cy7) as fluorescent markers can create up to 24 different colors.

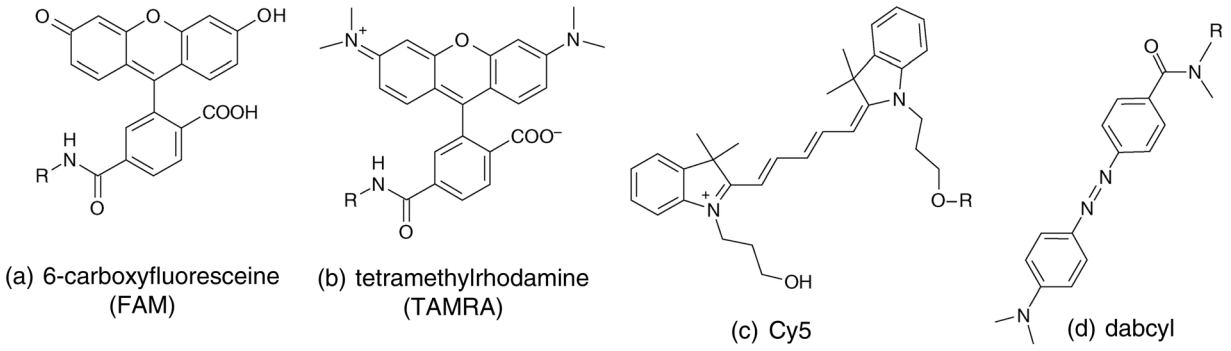


Figure 28.23 Examples of FRET components: (a) 6-carboxyfluorescein (FAM), (b) tetramethylrhodamine (TAMRA), (c) Cy5, and (d) dabcyl.

28.5 Amplification Systems

Often, detection is coupled with nucleic acid amplification procedures. Three types of amplification formats are known:

- *target amplification:* amplification of the nucleic acid to be detected;
- *target-specific signal amplification:* signal amplification coupled with target hybridization;
- *signal amplification:* amplification of the signaling components.

Table 28.4 gives an overview of various amplification reactions. Target amplification procedures have many advantages. In most cases, they lead to an exponential amplification and

Table 28.4 Overview of amplification reactions. Source: Kessler, C. (1994) Non-radioactive analysis of biomolecules. *J. Biotechnol.*, **35**, 165–189; and Kessler, C. (ed.) (2000) *Non-radioactive Analysis of Biomolecules*, Springer, Berlin, Heidelberg.

Type of amplification	Examples
Target amplification	
Replication	
Temperature cycles	PCR: polymerase chain reaction
cDNA synthesis/temperature cycles	RT-PCR: PCR connected with cDNA synthesis
Isothermal cycles	PCR: *in situ* PCR
Transcription	
Cyclical isothermal cDNA synthesis	NASBA: nucleic acid sequence based amplification
	TMA: transcription mediated amplification
	TAS: transcription based amplification system
	3SR: self-sustaining sequence replication
Increased rRNA copy number	16S/23S rRNA probes
Target specific signal amplification	
Replication	
Isothermal replication cycles	Qβ replication
Ligation	
Temperature cycles	LCR: ligase chain reaction
Replication/ligation	
Temperature cycles	RCR: repair chain reaction
Probe hydrolysis	
Isothermal cyclic RNA hydrolysis	CP: cycling probes
Displacement of indicator probes on flap structures by cleavase enzymes and amplification of indicator probes	Invader technology
Signal of amplification	
Tree structure	
Probe network	Hybridization trees, branched probes
Network of indicator molecules	PAP: peroxidase: anti-peroxidase complex APAAP: alkaline phosphatase:anti-alkaline Phosphatase complex
Enzyme catalysis	
Enzymatic substrate turnover	ELISA: enzyme linked immunosorbent assay
Polyenzyme	Enzyme gel conjugate
Coupled signal cascades	
Cyclic $NAD^+/NADP^+$ redox reaction	*Self* redox cycle

therefore to high amplification rates (10^6–10^9). In addition, the complexity of the detection system is significantly decreased, since only the targets to be detected are amplified and not unspecific sequences. An example of a target amplification is the polymerase chain reaction (PCR), in which the section of nucleic acid to be detected is enzymatically reproduced in a primer-dependent reaction.

Polymerase Chain Reaction, Chapter 29

This reduction in complexity is absent in plain signal amplifications; consequently, for complex genomic DNA, like the human genome, they are only used in combination with ceding PCR target amplification procedures. An example of a signal amplification combined with target amplification is the enzymatically catalyzed conversion of dyes or luminescence-generating substrates (ELISA).

ELISA, Section 5.3.3

In addition, plain signal amplification only leads to linear signal amplification and therefore lower amplification rates (10–10^3); the result is lower sensitivity of the detection reaction. Since unspecific hybridizations are also amplified, this often results in a less favorable signal-to-noise ratio, which decreases the specific detection of the target sequence.

The sensitivity reached in systems without target amplification, but with ELISA signal amplification, lies in the range of picograms (10^{-12} g) to femtograms (10^{-15} g). In combination with target amplification systems like PCR, the sensitivity goes up to the attogram (10^{-18} g) range. Such reactions allow the detection of single molecules. In this range the sensitivity reached in practice is no longer limited by the sensitivity of the detection system, instead it is limited by statistical effects relating to preparation of the test sample.

While signal amplification is an important component of protein or glycan analysis, only nucleic acid analysis allows amplification of the target. This enables the combination of both types of amplification for the detection of nucleic acids, which leads to very high detection sensitivities.

28.5.1 Target Amplification

Target Elongation Target elongations are thermocyclic reactions in which both strands of the DNA to be detected are replicated. The most important method for target elongation is PCR. RNA must first be transcribed into cDNA with reverse transcriptase (RT PCR) prior to use in a PCR. In homogeneous detection systems, PCR amplification is often carried out in *TaqMan* or LightCycler format and fluorescence detection is employed. Strand displacement amplification (SDA) is another means to amplify DNA, in this case under isothermal conditions.

In the case of *in situ* PCR, these amplification reactions can be carried out in fixed tissues or cells for the specific amplification of target sequence fragments.

Target Transcription The amplification reactions involved in transcription are isothermal and cyclic through reverse transcription and transcription of an intermediately formed transcription unit. The starting material is target RNA, which is converted into cDNA with the help of promoter-containing primers, forming intermediate double-stranded DNA molecules containing a T7, T3, or SP6 promoter. These intermediary transcription units are transcribed into the starting RNA. There are different ways to carry out the transcription amplification. The reactions differ, among other things, in that only one or both primers carry promoter sequences. One variant, nucleic acid sequence-based amplification (NASBA), is described in the following section. An important alternative is transcription-mediated amplification (TMA).

Nucleic Acid Sequence-Based Amplification (NASBA), Section 29.6.1

***In Vivo* Amplification** Special systems for the detection of bacteria exploit the fact that ribosomal RNA includes species-specific sequences and *in vivo* are present at a copy number of 10^3 to 10^4, so they can be easily detected. The indicator sequences are the variable regions of the rRNA or intergenic spacer regions, i.e. segments between the rRNA genes.

28.5.2 Target-Specific Signal Amplification

This type of reaction does not involve the amplification of the target itself, instead a nucleic acid tag or oligonucleotide, which hybridizes with the target, is replicated or changed. These reactions have several disadvantages relative to direct amplification of the target: For one thing, it involves a pure detection reaction, since no new target sequences are created. For another, these detection reactions have a limited specificity, since a target-coupled signal is increased and therefore unspecific amplification products cannot be filtered out as, for example, they would be with a subsequent hybridization with a target-specific probe.

An example for target-specific signal amplification is the ligase chain reaction (LCR), which will be briefly described in the next chapter. Another example is the cleavase reaction, which

Ligase Chain Reaction (LCR), Section 29.6.4

involves cleavage of a mismatched end of a probe, a flap, by the enzyme cleavase, which recognizes and cuts such ends. This leaves an extra base on the end of the cleaved free end of the oligonucleotide, which then allows the amplification of a signal sequence after binding to a complementary indicator probe, which allows elongation of its 3′ end.

28.5.3 Signal Amplification

This sort of amplification includes reactions to boost the resulting signal, instead of the number of targets, and brings only limited increases in sensitivity (10–10^3). Unspecific reactions mentioned for the target-independent amplifications can lead to unspecific signals. Nevertheless, the signal amplifications are of great value, particularly when combined with target amplification reactions.

Branched Structures Branched structures are built up from forked probes, which contain target-specific sequences as well as sequences for the generation of signal (Christmas trees, probe brushes). The branched probes are complemented by universal detection probes, which are linked to a marker enzyme, such as AP. The primary, bivalent probes are composed of sequences that have specific sequences complementary to the branched secondary probes. The secondary branched probes are composed of sequences complementary to the primary probes as well as branched structures complementary to a third probe. The third components are universal, enzyme-labeled AP probes. By hybridization between the target, primary (bivalent), secondary (branched), and tertiary (AP-labeled) probes, complex structures are built up that lead to amplification of the signal. An example is the use of branched DNA (bDNA) for the identification of hepatitis C virus (Figure 28.24). Additional sensitivity can be achieved through the cassette-like attachment of several branched structures to a target. A disadvantage of this system is, besides the limited amplification rates, the complex set up and therefore the limited ability to control the branched structures, as well as the possibility of the occurrence of unspecific hybridization. These can lead to unspecific background reactions and therefore to a poor signal-to-noise relationship, which limits its sensitivity.

Enzyme Catalysis The amplification of signals through the enzymatic conversion of substrate with the help of marker enzymes is used in direct and indirect detection systems (e.g., AP, POD, β-Gal; Section 28.4.3). Enzyme catalysis as part of the detection reaction

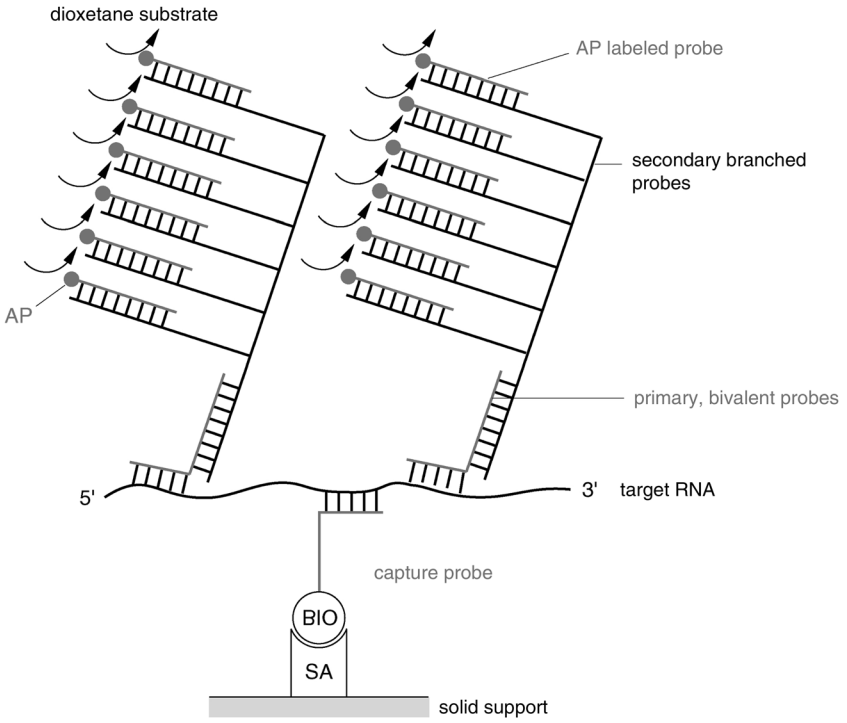


Figure 28.24 Branched structures for signal amplification. The structural components are the capture probe, the bivalent primary detection probe, the branched secondary probes, and the universal AP detection probes. The capture probe can also be covalently bound directly to a membrane.

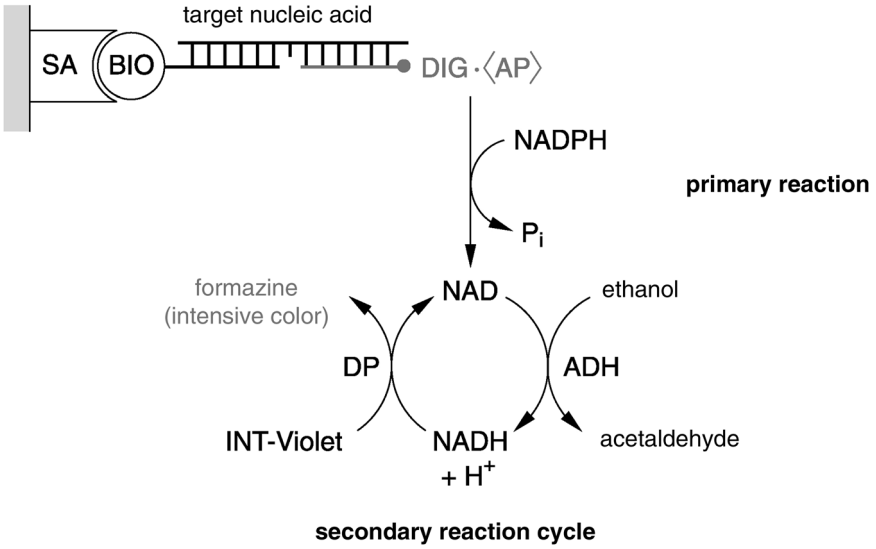


Figure 28.25 Signal amplification by coupled cyclic ADH: diaphorase redox reaction. ADH: alcohol dehydrogenase; DP: diaphorase.

leads, depending on the reaction format and type of marker system, to up to a 10^3-fold increase in sensitivity. The use of polyenzymes leads to an additional increase in sensitivity by a factor of 3–5.

In this context, coupling of PCR and enzyme catalysis with luminescent substrates is often the best choice. For example, the DIG hapten (Section 28.4.3) is often incorporated into the amplicon by the primers or nucleotides during PCR; the DIG-marked amplicon is subsequently detected with high sensitivity with the help of AP-antibody conjugates by the catalytic conversion of indolyl or dioxetane substrates. By the combination of amplification types, sensitivities of up to the detection of single molecules are achieved.

Coupled Signal Cascades For this reaction type, a primary substrate conversion is coupled to a secondary enzymatic reaction. If the secondary reaction is cyclic, a signal cascade results. An example of a signal cascade is the self-redox cycle shown in Figure 28.25.

The primary reaction is catalyzed by the marker enzyme alkaline phosphatase (AP) that is linked, either directly or indirectly, with the hybridization probe. Primary NADP is dephosphorylated by AP to NAD. The formed NAD activates a secondary redox cycle, in which it is reduced to NADH by the enzyme alcohol dehydrogenase (ADH), coupled with the oxidation of ethanol to acetaldehyde. The secondary reaction cycle is completed by the enzyme diaphorase (DP), which re-oxidizes the NADH to the original NAD in a further coupled reaction and at the same time reduces the dye NBT-violet to deeply colored formazine. Through the enzyme diaphorase, the secondary reaction cycle is closed.

The soluble formazine can be quantitated photometrically ($\lambda = 465$ nm); the signal amplification leads to a 10- to 100-fold increase in sensitivity or, alternatively, to much shorter reaction times. Traces of NAD contamination in the primary substrate NADPH, which accumulate during extended storage, generate a background, which can critically compromise the sensitivity.

Further Reading

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., and Struhl, A. (1987–2005) *Current Protocols in Molecular Biology*, vol. 1–4, Suppl. 1–69, Greene Publishing Associates and Wiley-Interscience, New York.

Clarke, J.R. (2002) Molecular diagnosis of HIV. *Expert Rev. Mol. Diagn.* **2**, 233–239.

Griulietti, A., Overbergh, L., Valckx, D., Decallonne, B., Bouillon, R., and Mathieu, C. (2001) An overview of real-time quantitative PCR: applications to quantify cytokine gene expression. *Methods*, **25**, 386–401.

Hares, B.D. and Higgins, S.J. (1995) *Gene Probes*, vol. 1 and 2, IRL Press, Oxford.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
Keller, G.H. and Manak, M.M. (1993) *DNA Probes*, 2nd edn, Stockton Press, New York.
Kessler, C. (1991) The digoxigenin (DIG) technology – a survey on the concept and realization of a novel bioanalytical indicator system. *Mol. Cell. Probes*, **5**, 161–205.
Kessler, C. (ed.) (1992) *Non-radioactive Labeling and Detection of Biomolecules*, Springer, Berlin, Heidelberg.
Kessler, C. (ed.) (2000) *Non-radioactive Analysis of Biomolecules*, Springer, Berlin, Heidelberg.
Kidd, J.M. *et al*. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64. See also http://www.hgsc.bcm.tmc.edu/
Kricka, L.J. (1995) *Nonisotopic Probing, Blotting, and Sequencing*, Academic Press, San Diego.
Kricka L.J. (2002) Stains, labels and detection strategies for nucleic acids assays. *Ann. Clin. Biochem,*. **39**, 114–129.
Lander, E.S., Linton, L.M., and Birren, B. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921 and 411, 720 and 412, 565.
Lee, H., Morse, S., and Olsvik, O. (1997) *Nucleic Acid Amplification Technologies*, Eaton Publishing, Natick.
Marshall, A. and Hodgson, J. (1998) DNA chips: an array of possibilities. *Nat. Biotechnol.*, **16**, 27–31.
McPherson, M.J., Quirke, E., and Taylor, G.R. (1996) *PCR – A Practical Approach*, vol. 1 and 2, IRL Press, Oxford.
Nuovo, G.J. (1997) *PCR In Situ Hybridization: Protocols and Applications*, 3rd edn, Lippincott-Raven Press.
Nussbaum, R.L., McInnes, R.R., and Willard, H.F. (2001) *Thompson & Thompson Genetics In Medicine*, 6th edn, W.B. Saunders, Philadelphia.
Ørum, H., Kessler, C., and Koch, T. (1997) Peptide nucleic acid, in: Lee, H., Morse, S., and Olsvik, O. (eds) *(1997) Nucleic Acid Amplification Technologies*, Eaton Publishing, Natick.
Persing, D.H., Smith, T.F., Tenover, E.C., and White, T.J. (1993) *Diagnostic Molecular Microbiology: Principles and Applications*, American Society for Microbiology, Washington.
Ramsay, G. (1998) DNA chips: state-of-the art. *Nat. Biotechnol.*, **16**, 40–44.
Reischl, U., Wittwer, C., and Cockerill, F. (eds) (2002) *Rapid Cycle Real-Time PCR – Methods and Applications*, Springer-Verlag, Berlin, Heidelberg.
Tyagi, S., Bratu, D., and Kramer, E.R. (1997) Multicolor molecular beacons for allele discrimination. *Nat. Biotechnol.*, **16**, 49–53.
Venter, J.C., Wittwer, C.T., and Herrmann, M.G. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351;
Venter, J.C., Wittwer, C.T., and Herrmann, M.G. *et al*. (2001) *Science*, **292**, 1838.

Polymerase Chain Reaction

29

Nancy Schönbrunner,[1] Joachim Engels,[2] and Christoph Kessler[3]

[1]Roche Molecular Systems, Inc., 4300 Hacienda Dr., Pleasanton, CA 94588, USA
[2]Goethe University Frankfurt, Institute of Organic Chemistry and Chemical Biology, Department of Biochemistry, Chemistry and Pharmacy, Max-von-Laue Straße 7, 60438 Frankfurt am Main, Germany
[3]PD Christoph Kessler, Consult GmbH, Icking, Schloßbergweg 11, 82057 Icking-Dorfen, Germany

The polymerase chain reaction (PCR), a method for the amplification of nucleic acids, is one of the greatest scientific discoveries of the recent past. Without exaggeration it can be said that this discipline, though still young, has revolutionized molecular biology. The possibilities of PCR appear to be almost unlimited. The number of articles in scientific journals about improvements, new applications, and breakthroughs in the areas of basic and applied research, as well as medicine, diagnostics, and other areas, grows daily. The history of its discovery, during a night drive through the mountains of California, is portrayed in an impressive article by Kary B. Mullis, its inventor. He writes: "The polymerase chain reaction was not the result of a long development process or a lot of experimental work. It was invented by chance on a late evening in May, 1983 by the driver of a gray Honda Civic during a drive along Highway 130 through the mountains between Cloverdale and Anderson Valley in California."

Since the discovery of PCR in 1983, around 800 000 publications in diverse journals and magazines have appeared as of March 2014.

PCR applications lie in the most diverse areas: molecular biological basic research, cloning of defined sequence segments, generation of samples, genetics, medicine, genome diagnostics, forensics, food sector, plant breeding, agriculture, environment, and archeology, to just name a few.

The PCR reaction has been integrated into the most diverse processes: PCR amplification with subsequent electrophoretic separation of the amplification products, cloning and sequencing of the amplified sequences, allelic PCR for the elucidation of mutations, analysis of the PCR products in blot detection formats, coupling with quantitative heterogeneous and homogeneous detection methods, or *in situ* PCR for the amplification of particular target sequences directly in tissues or cell cultures. For biological samples an important requirement is for the upfront sample preparation to extract and concentrate nucleic acid from the biological matrix and remove inhibitors.

Alternatives to PCR have also been established, with NASBA (nucleic acid sequence-based amplification) and TMA (transcription-mediated amplification) amplification procedures being the best known. While PCR begins with DNA as the starting nucleic acid and the amplification is accomplished by temperature cycles, NASBA and TMA amplifications begin with RNA as the starting nucleic acid and the amplification cycles are carried out at a constant temperature, such as 42 °C.

29.1 Possibilities of PCR

The PCR method as a means to amplify any given nucleic acid segment relies on an idea as brilliant as it is simple. To estimate the value of such an amplification, we will begin by considering conventional analysis procedures. For example, gel electrophoresis has a lower

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

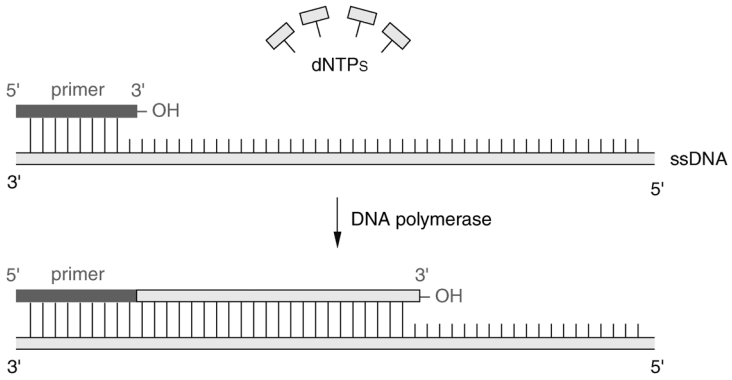


Figure 29.1 Schematic of the polymerization of DNA. In the presence of a primer with free 3′-OH ends and free nucleotides (dNTPs) DNA polymerases convert single-stranded (ssDNA) into double-stranded DNA.

limit of detection of about 5 ng DNA. If you calculate the number of molecules it ends up being 10^{10} for a fragment 500 base pairs (bp) in length. To increase the sensitivity, the DNA in such gels can be transferred to solid supports in order to detect them with radioactive or non-radioactive probes. This increases the sensitivity such that approximately 10^8 molecules can be detected; however, for many diagnostic purposes this is not even close to being sufficient. In viral diagnostics titers of under 1000 particles per milliliter blood are often present. Even the most sensitive current analysis procedures do not come close to reaching the sensitivity of PCR: it is theoretically capable, under optimal conditions, of generating up to 10^{12} identical molecules from a single nucleic acid segment in a few hours, which then are available for diagnostics or other analytical method (Section 29.5).

How is this accomplished? PCR takes advantage of the ability of DNA polymerases to duplicate DNA. A requirement for this is a short segment of double stranded DNA with a free 3′ OH end, which can correspondingly be extended (Figure 29.1). Mullis realized that such a short segment can be created artificially by adding DNA fragments of about 20 nucleotides in length, also referred to as oligonucleotides or primers. These bind, or anneal, to the ends of the DNA strand to be amplified and can now be extended by the polymerase. If the newly synthesized double strand DNA is denatured by increasing the temperature, new primer molecules can bind upon cooling and the process can begin again. If two primers are added to the sample, one of which binds to the sense strand, the other the antisense strand, after each cycle of new synthesis and denaturation a doubling of the segment between the primers takes place. PCR leads to an exponential amplification, since the new strands are available as templates for the next round of amplification (Figure 29.2). And something else was recognized by Mullis's group: if one uses a temperature stable DNA polymerase, such as those found in organisms that live in hot springs, it is possible to run the reaction without interruption.

29.2 Basics

29.2.1 Instruments

The required reagents and tools for PCR are quite simple. In the course of time they have been steadily improved in terms of data security, throughput, and user comfort. The first thermocyclers consisted of three water baths set to different temperatures and the samples were moved by hand and with the aid of a stopwatch from one bath to the next. Later, robotic arms took over this task. Today, they are relatively compact devices that hold the PCR samples in a 96-well plastic micro-volume plate in a metal block, which is systematically heated and cooled. A significant difference of modern thermocyclers is their heating technology, which employs either Peltier elements or works with the aid of liquids. The newest developments in this field are aimed at a drastic increase in speed through miniaturization – PCR in a glass capillary tube with a very low volume (*LightCycler*®) or the amplification and detection in a single device (PCR combined with simultaneous FRET detection (*TaqMan*®)). Devices of this sort allow real time detection of a PCR product during the amplification (Section 29.7).

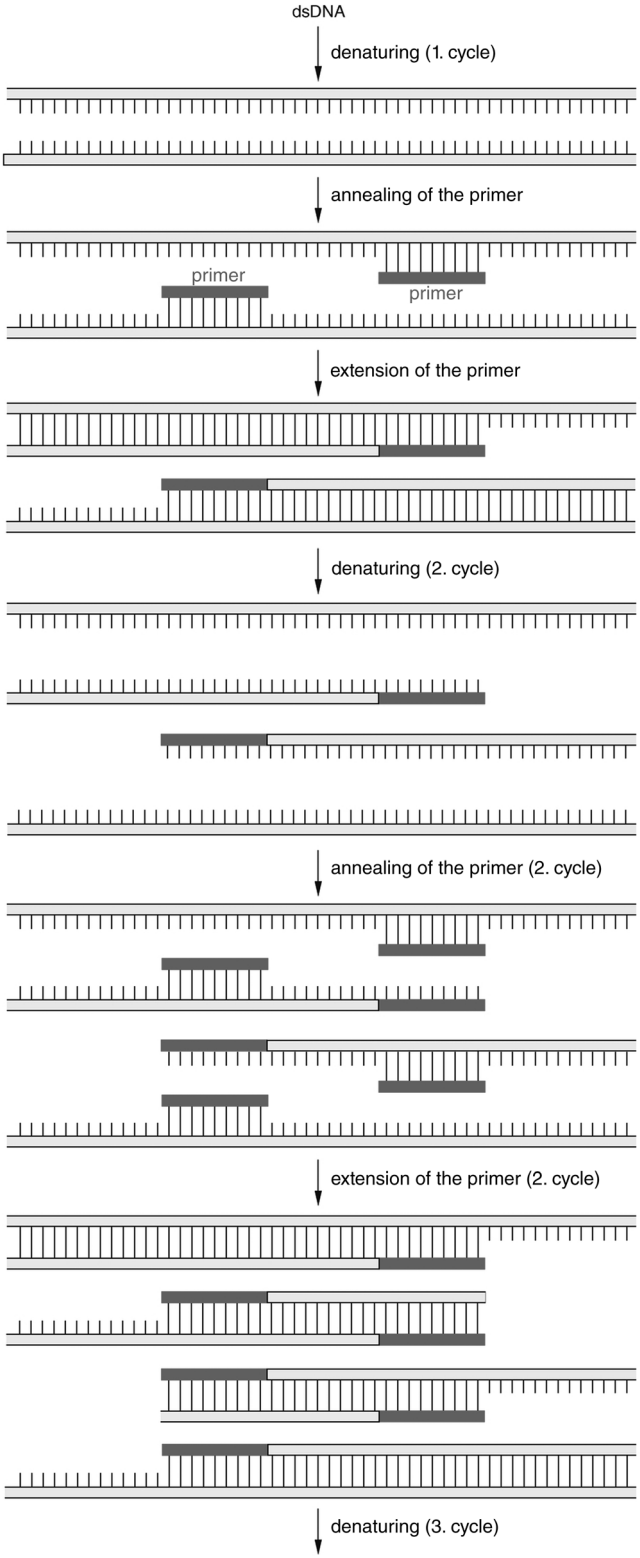


Figure 29.2 Schematic of PCR. The number of DNA segments can theoretically double in each cycle of a process of denaturation, primer annealing, and primer extension. The first two cycles are shown. The number of the copies grows exponentially with each round.

29.2.2 Amplification of DNA

Probe Preparation If the starting material for PCR comes from biological material, the nucleic acids must first be released from, for example, virus particles, bacteria, or cells and separated from interfering components like proteins, lipids, or inhibitors, such as hemoglobin degradation products in blood. Probe preparation, besides releasing and purifying the samples, leads to the concentration of the nucleic acids.

Various methods are used: A common method uses the property that DNA binds to glass in the presence of chaotropic salts. If the glass particles have a magnetic core, they can be captured by placing a magnet on the vessel wall, separated from interfering substances by wash steps, and the surface bound nucleic acids eluted and concentrated into a small volume of solution by suitable elution buffer. The advantage of this method is that it is generic, thus it can be used on different target sequences. Another method uses hydroxyapatite columns. If the goal is to purify particular target sequences for PCR, capture beads, which have capture probes on their surface, can be used. The capture probe binds to the particle by the binding pair streptavidin (adsorption to the particle surface) and biotin (tagging the capture probe).

Complementary target sequences bind to the capture probe, and if the particles employed are also magnetic then separation from interfering substances can take place by applying a magnetic field.

Cycles A typical PCR run consists, as a rule, of three stages at different temperatures. This is particularly easy to visualize as a temperature–time profile (Figure 29.3). The reaction is started by heating to 92–98 °C. This step serves to denature DNA into its single strands. Since the initial DNA is present in a complex, high molecular weight structure, a time of 5–10 min is chosen to ensure that even GC-rich sequences are denatured. The second step of the reaction is annealing of the primer. For this to take place, the reaction must be cooled to a probe-specific temperature. Annealing of the primer to a single strand of the target sequence critically influences the specificity of the PCR. The annealing has a critical influence on the specificity of the PCR. After the annealing step the temperature is increased to 72 °C, the optimum temperature of the enzyme used, *Taq* DNA polymerase. For both the annealing and extension steps a time of less than a minute is usually sufficient. Only for very long PCR products of over a kilobase is the extension time lengthened in order to be sure that the complete strand is synthesized. This is important, since only completely extended DNA strands can function as templates in the next cycle of PCR. In the next step of the reaction it is again heated to 92–95 °C, in order to separate the product double-stranded DNA into single strands. Since in the ideal case only the newly synthesized segments are

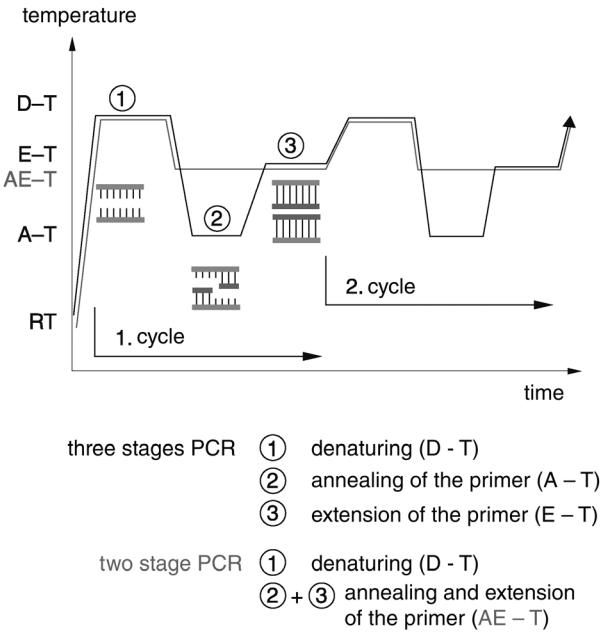


Figure 29.3 Temperature/time profile of PCR (two and three step).

Table 29.1 Summary of PCR Master mixes (100 µl end volume).

Reagent	Final concentration
Taq DNA polymerase	2–5 units 1×
10 × *Taq*-Buffer (100 mM Tris/HCl pH 8.3; 500 mM KCl)	1×
10 mM nucleotide mixture (dATP, dCTP, dGTP, dTTP)	0.2 mM
$MgCl_2$	0.5–2.5 mM
Primer I	0.1–1 µM
Primer II	0.1–1 µM
H_2O	Variable
Templates	Variable

present as double-strands, after the second cycle a much shorter denaturation time of 10–60 s is necessary. Newer protocols combine annealing and extension into a single step, usually at 62–72 °C. This makes a two-step PCR out of a three step one.

For most applications, enough product for further analysis is present after 30–35 cycles. Amplification only requires 40–50 cycles in exceptional situations or for nested PCR (Section 29.3.1).

Enzyme The most important requirements for DNA polymerases in PCR are to have a high processivity, the ability to synthesize long stretches of DNA, and/or rapid binding and extension kinetics at 72 °C, as well as a very high temperature stability at 95 °C. The polymerases that have these characteristics are *Taq*, *Tth*, *Pwo,* and *Pfu* DNA polymerases. *Taq* DNA polymerase is the most commonly used in standard protocols. *Tth* DNA polymerase, like *Taq* DNA polymerase, has a high 5′-3′ polymerization activity, but in addition *Tth* possesses reverse transcriptase activity under certain conditions. This is explained in more detail in Section 29.2.3. *Pwo* and *Pfu* DNA polymerases have, in addition to their polymerization activity, a 3′ exonuclease activity. This activity is referred to as proofreading activity, since these enzymes are capable of recognizing and removing an incorrectly incorporated nucleotide so that the mistake can be corrected with a new round of polymerization. Besides the single polymerases, mixtures of the enzymes are often offered commercially. Examples of the reagents needed for amplification according to standard protocols are listed in Table 29.1.

Buffer The buffer conditions need to be set according to the requirements of the polymerase. The ion concentration of the buffer, usually provided with the enzyme, is very important, since it influences the specificity and processivity of the total reaction. The buffers usually provided with *Taq* DNA polymerase usually come with and without magnesium chloride. For the optimization of a new PCR (Section 29.2.4), buffers without magnesium chloride are better, since the range of conditions is much greater with an additional magnesium chloride solution. Other possible additives are bovine serum albumin (BSA), Tween 20, gelatin, glycerol, formamide, and DMSO. This can lead, in some cases, to stabilization of the enzyme and to optimized primer annealing.

Nucleotides The concentration of the four deoxynucleotide triphosphates (dATP, dCTP, dGTP, dTTP) is usually in the range 0.1–0.3 µM. All four dNTPs should be present in an equimolar amount. The only exception is when other nucleotides are used for amplification, for example dUTP. Nucleotide analogs are often used in excess, mostly 3 : 1, since they are not incorporated as well by *Taq* DNA polymerase.

Primers Important prerequisites for an optimal PCR, measured on the basis of specificity and sensitivity, is the selection of the primers. There are four basic types of primers:

- sequence-specific primers,
- degenerate primers,
- oligo(dT) primers (only for RNA),
- short random primers (usually for RNA).

Oligo(dT) primers and random primers (hexanucleotides) are usually used for the amplification of RNA. They are described in the following section. The use of degenerate primers is limited to particular questions and will be treated separately in Section 29.3.3. The primers used most often, by far, are sequence-specific primers. To ensure that they actually bind to their target sequences, there are several rules that must be followed. Some of the most important are mentioned briefly here.

Primer requirements

1. at least 17 nucleotides long (usually 17–28 nt);
2. balanced G/C to A/T content;
3. melting point between 55 and 80 °C;
4. melting point of the forward and reverse primers should be as close as possible;
5. no hairpin structures, particularly on the 3′-end (Figure 29.4);
6. no dimerization: neither by itself nor with the second primer (Figure 29.4);
7. no G/C nucleotide on the 3′-end, if possible, since this increases the danger of mispriming;
8. no "strange" base sequences like poly(A), more than four Gs, or long G/C stretches, if possible.

Today diverse computer programs support the user in the search for suitable primer sequences and, in addition, show their melting points. This can be calculated with various formulas. The simplest of these formulas calculates a temperature of 2 °C for each A or T and for every G or C a temperature of 4 °C. For a primer of 20 nucleotides in length, a 20-mer, with a balance number of A/T and G/C, it calculates a melting point of 60 °C.

Specificity of the Hybridization and Stringency, Section 28.1.2

Templates (Genomic DNA, Plasmids, Viral DNA) The most important influence of the target on the success of PCR is the length of segment to be amplified, the sequence of the primer binding sites, and the number of input molecules. A microgram of human genomic DNA contains 3×10^5 target sequences, provided it is a single copy gene and not a repetitive element. The same mass of a plasmid of 3 kb DNA contains 3×10^{11} molecules. In other words 1 µg genomic DNA contains as many molecules as 1 pg of plasmid DNA. This needs to be taken into account when using different templates, particularly when used for preparative purposes.

The maximum amplification length of a DNA is primarily determined by the processivity of the polymerase used. Today there are enzymes and enzyme mixtures that allow amplification of fragments up to 40 kb. In such cases the extension time must be much longer, up to 30 min per cycle. In general, short segments of 0.1–1 kb in length are favored, since these can be optimally amplified with PCR.

Besides the length and the number of molecules, the primer binding sites also determine whether PCR is successful. To avoid mispriming, repetitive sequences should be excluded and instead a single copy site should be selected for the primers.

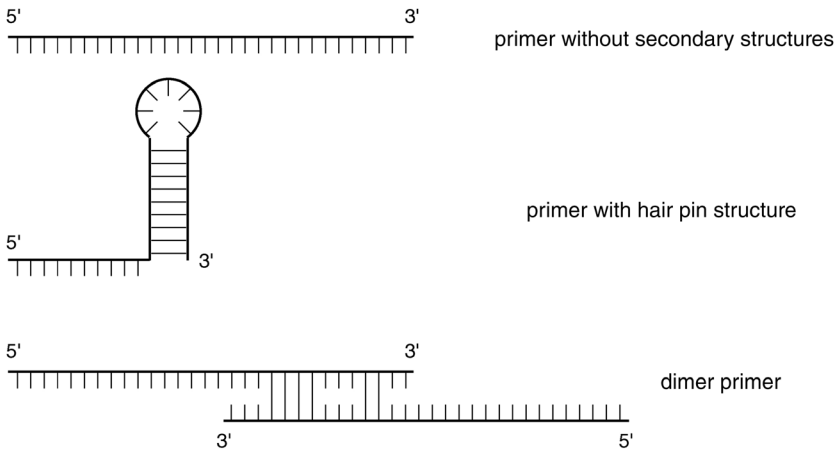


Figure 29.4 Secondary structures of primers. For the selection of primers it is important to avoid secondary structures. Complementary segments between the sense and antisense primers must also be factored into the analysis.

29.2.3 Amplification of RNA (RT-PCR)

Many methods to analyze RNA exist in molecular biology, such as Northern blots, *in situ* hybridization, RNAse protection assays, and nuclease S1 analysis, to name just a few. All these methods have the disadvantage, however, that they are time consuming and often not sensitive enough. This is particularly true for the analysis of low copy number transcripts or for viral RNA, which is only present at very low starting concentrations. Adaptation of PCR technology to allow the amplification of RNA led to many new discoveries and more sensitive diagnostics. It can be used to investigate gene expression in cells or, with the aid of quantitative RT-PCR, to determine the amount of a specific mRNA or viral RNA. In addition, with oligo(dT) priming (see below), complete cDNA banks can be created, which enables an overview of tissue-specific expression.

In situ Hybridization, Section 35.1.4
Ribonuclease Protection Assay (RPA), Section 34.1.3
Nuclease S1 Analysis of RNA, Section 34.1.2

Enzymes Since the starting RNA cannot be directly used as a temple by *Taq* DNA polymerase, the RNA must first be transcribed into DNA to be amplified. There are several enzymes, called reverse transcriptases (RTases) or RNA-dependent DNA polymerases, for this purpose. The newly synthesized strand is termed complementary DNA (cDNA) and the step in which this cDNA is created is called reverse transcription (RT). The complete reaction of RT and amplification is therefore called RT-PCR (Figure 29.5). Several different reverse transcriptases can be used for this purpose.

MMLV RTase The enzyme comes from Moloney murine leukemia virus, has an optimum temperature of 37 °C, and is able to synthesize cDNA up to a length of 10 kb due to its high processivity. The optimum pH is 8.3.

AMV RTase Isolated from avian myeloblastosis virus (AMV) from birds, its optimum temperature is 42 °C and it has a similarly high processivity as MMLV RTase. Its optimum pH is 7.0.

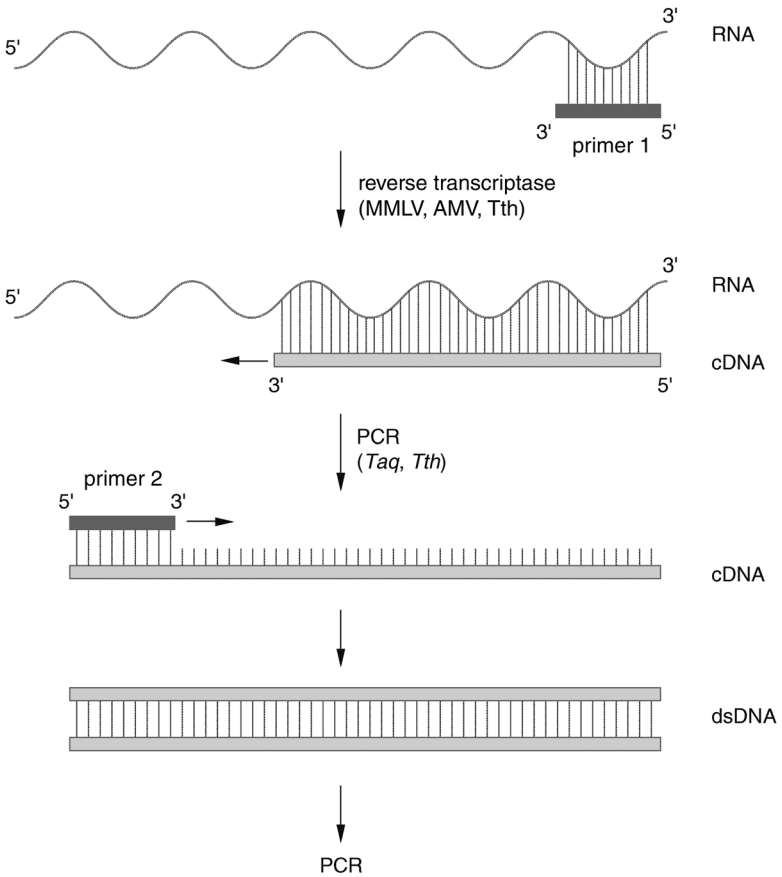


Figure 29.5 Schematic portrayal of RT-PCR. Since RNA cannot be amplified directly by PCR, it must first be transcribed into cDNA. Enzymes that catalyze this step are MMLV RTase, AMV RTase, and *Tth* polymerase.

***Tth* DNA Polymerase** This heat stable enzyme comes from the bacteria *Thermus thermophilus*. In contrast to the other two enzymes, *Tth* DNA polymerase possesses two activities: In the presence of manganese ions it has both RT and DNA polymerase activity. Since the *Tth* DNA polymerase comes from a thermophile, just like *Taq* polymerase, it has a temperature optimum of 60–70 °C. It is the only enzyme capable of carrying out both steps of an RT-PCR under the same buffer conditions. For the RT step a high concentration of manganese ions is optimal, but it tends to inhibit DNA polymerase. This affects the processivity of the enzyme. For this reason, *Tth* DNA polymerase is only able to synthesize 1–2 kb of cDNA.

Procedure After using different enzymes, there are also different possibilities to carry out the reaction for a RT-PCR.

In Two Reaction Tubes The first RT step is carried out in a relatively small volume (to increase the sensitivity). This has the advantage that the reaction conditions can be set optimally for the RT enzyme used. After the RT step, the entire reaction, or an aliquot, is removed and a "normal" PCR is carried out. The disadvantage of this procedure is an increased danger of contamination (Section 29.4), since an additional pipetting step is required. For such a two-step process, AMV or MMLV RTase is used for the RT and *Taq* DNA polymerase for the PCR.

In a Single Reaction Tube For the reason mentioned above, it is advantageous to carry out the whole RT-PCR reaction in one tube without the need to pipette from one tube into another. This is possible, in principle, with all three reverse transcriptases, partially in combination with *Taq* polymerase; however, *Tth* DNA polymerase is particularly well suited. In almost all cases, the lower processivity is an acceptable price to pay, since the length of the segment to be amplified is less than 2 kb. The main advantage, however, of using *Tth* DNA polymerase is the possible increase of the reaction temperature of the RT step to 60 °C. This temperature helps to resolve secondary structures and thereby eases annealing of the primer.

Primer For RT-PCR, three different types of primers can be used (Figure 29.6):

- Sequence-specific primers anneal specifically in both the RT step and the following amplification to the same site of the RNA or cDNA. These are most often used for diagnostic tests for the detection of viral RNA.
- Oligo(dT) primers consist of a segment of 12–18 dTs, which semi-specifically bind to the polyA tail of eukaryotic mRNAs. They are only useful for the RT step. For the following amplification steps further, sequence-specific primers are needed.
- Short random primers are a mixture of hexanucleotides of different sequences. They bind "randomly" to the RNA and lead to a pool of cDNAs of various lengths, which, like the oligo (dT) primers, are then amplified with a second set of sequence-specific primers.

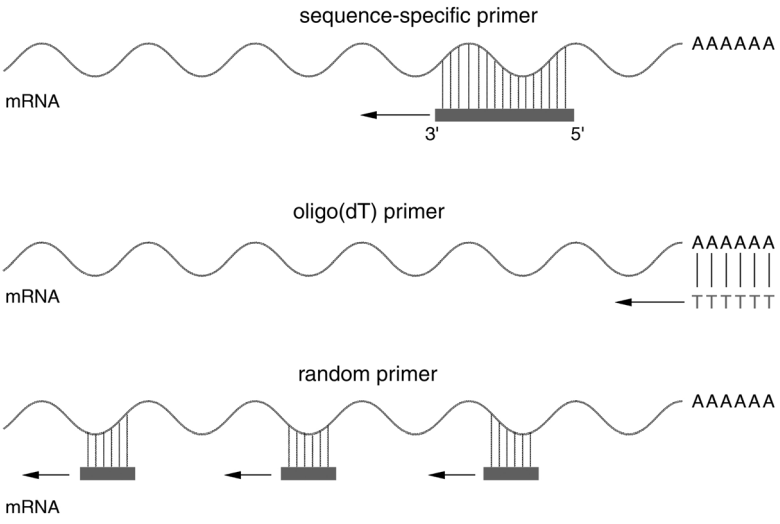


Figure 29.6 Different priming methods in RT-PCR.

29.2.4 Optimizing the Reaction

One of the most important and time-consuming parts of the establishment of a new PCR is optimization of the reaction. Among the analytical aspects, the sensitivity is particularly important. This is critical in forensic medicine or the detection of very low amounts of DNA or RNA in infectious disease. In cases where the amplification is for preparative purposes, such as by the synthesis of probes or templates for sequencing, the yield, or amount of the PCR product formed, is of primary importance. This section will explain a few of the important adjustments that can be made and a few strategies for the optimization.

DNA Amplification

Choice of the Primers The selection of primers requires a substantial investment of time. If the sequence segments allow, several primers pairs should be tested, since secondary structure of the target sequence is, in principle, always to be expected. If no PCR product is synthesized, it can be helpful to successively lower the annealing temperature. However, one must pay particular attention to the possibility of nonspecific amplification.

Magnesium Ions An important factor, which determines the processivity and overall activity of *Taq* polymerase, is the concentration of magnesium ions. In the first experiments a concentration of between 0.5 and 5 mM should be explored.

Additives Many additives to the reaction solution can help stabilize *Taq* polymerase or annealing of the primers. These include glycerol, BSA, and PEG. Denaturation is aided by the addition of DMSO, formamide, Tween 20 Triton X-100, or NP40. Detergents may also stabilize the DNA polymerase.

Hot Start PCR If nonspecific amplifications are a problem, in some cases a hot start PCR can help: To suppress polymerization of nonspecifically hybridized primers at low temperatures, the activity of *Taq* polymerase is controlled such that it only begins at higher temperatures. One possibility is to add the enzyme only after the sample has been heated. There are also commercially available antibodies against *Taq* polymerase that only denature at higher temperatures and set the enzyme free. There are also chemically modified versions of *Taq* DNA polymerase (*Taq* GOLD), which is inactive below 60 °C; the chemical adducts are hydrolyzed under the specific buffer conditions at elevated temperatures. Finally, some DNA polymerases are supplied with specific aptamers that inhibit activity below the melting temperature of the aptamer.

Templates The quality of the template is also very important for a successful PCR. Sample preparation should ensure that PCR inhibitors are efficiently removed. Of particular significance are degradation products of hemoglobin, for blood preparations, and ethanol, which is often used to precipitate DNA.

RNA Amplification In addition to the points mentioned above, there are a few more aspects to factor in for RT-PCR: Single-stranded RNA often has more secondary structure than DNA. Since the formation of secondary structure is still a very complex and poorly understood process, the current computer programs are only of limited use in this aspect of primer selection. Even in cases where primer design appears optimal, the synthesis of the PCR product may fail due to secondary structures that either prevent the annealing of the primers or block extension.

In this context, the use of *Tth* DNA polymerase has often proven advantageous, since the RT reaction takes place at 60 °C, which helps to melt such secondary structures.

In some cases, adding an RNase inhibitor is recommended, since RNA is fundamentally a much more vulnerable template than DNA.

RT-PCR is principally less efficient than PCR. Even under optimal conditions only about 10–30% of the RNA present is transcribed into cDNA, which is then available for further amplification.

29.2.5 Quantitative PCR

The quantification of nucleic acids with PCR or RT-PCR has become an essential component of diagnostic questions. This is particularly true for the diagnosis and monitoring of infectious diseases. Two examples of great significance are the AIDS-causing HIV (human

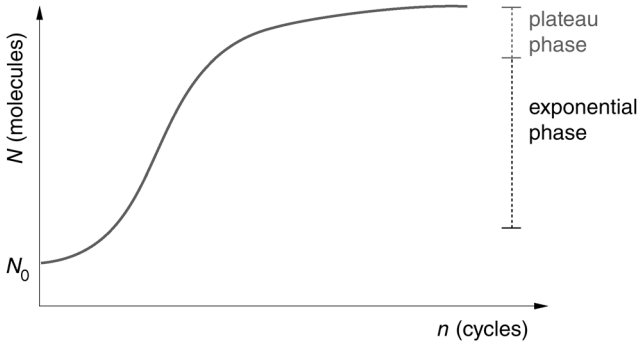


Figure 29.7 Typical course of a PCR. The kinetics of the PCR runs through an exponential phase into a plateau phase. The plateau comes about as a result of the build-up of inhibitors, competition between strand re-annealing with primer annealing, and because the reagents become limiting.

immunodeficiency virus) and the liver inflammation-causing hepatitis C-virus (HCV). In addition, in oncology there is interest in the quantitative measurement of mRNAs. The quantification is complicated by the fact that PCR is not a linear amplification. The exponential nature of the amplification means that small differences in reaction efficiency, such as in the presence of inhibitors in a particular sample, have profound effects on the yield of amplicon. The following equations demonstrate this:

$$N = N_0 2^n \tag{29.1}$$

where N is the number of amplified molecules, N_0 the number of molecules prior to amplification, and n the number of cycles. The number of molecules doubles under these (idealized) conditions with each cycle. In practice, however, the following formula applies:

$$N = N_0(1 + E)^n \tag{29.2}$$

where E is the efficiency of the reaction and has a value between 0 and 1. This value depends very strongly on the degree of optimization of the PCR. Experimentally, values for an optimized PCR have been found to be between $E = 0.8$ and $E = 0.9$.

The quantitative measurement is also made more difficult by the fact that towards the end of the amplification the exponential phase begins to plateau; this means that the value of E changes during the PCR (Figure 29.7). The maximum amount of product that can be generated during PCR is around 10^{13} molecules, but can deviate relatively strongly downwards.

Many methods for quantitative measurement have been developed and continuously improved over the years:

- *Limiting dilution:* A reference standard of known concentration is diluted in multiple steps, amplified, and the concentration is determined that was required in the preceding PCR to generate a barely detectable amplification. A sample to be measured is then also diluted in multiple steps to measure this same point. The number of dilutions then allows conclusions to be drawn about the starting concentration of the solution.
- *External standard:* The concentration of a sample to be measured is determined by comparing the signal it generated with that of a standard of known concentration

Both methods are not able to recognize internal interference in the amplification efficiency of an individual sample. The next generation of quantitative tests was focused on internal controls and standardized amplification reactions. There are two types of standardization:

- *Internal endogenous standard:* quantification of a so-called "house-keeping" gene.
- *Competitive (RT) PCR:* quantification using mimic fragments, which are added to the reaction and amplified along with the actual target sequence.

The following explains the last three options for quantification in more detail.

External Standard Samples of known concentrations are used to generate a curve to provide an external standard. The standard should be relatively similar to the intended

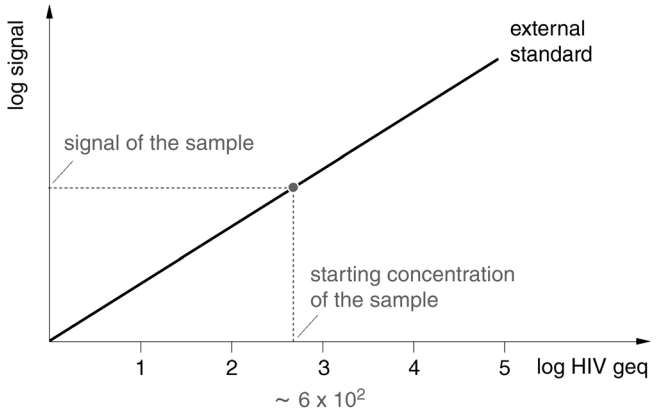


Figure 29.8 Quantification using an external standard. Shown is the measurement of a sample against a standard curve that was created with a known concentration of HIV from a cell line (geq, genome equivalents).

target sequence and should use the same primers for amplification. Very suitable, for example, is an HIV cell line that contains a known number of proviral genomes. This cell line is serially diluted, processed, amplified, and the signal plotted compared to the starting concentration (Figure 29.8). After amplifying the unknown sample, the signal is looked up on the standard curve to determine the starting concentration. The disadvantage is, as mentioned, the lack of an internal control of the reaction, ensuring that it ran properly and like the other samples. It is easy to imagine that even a low level of inhibition can lead to dramatic under-quantitation.

Internal Standardization DNA amounts in the samples are calibrated to internal sequences of the genome to provide internal standardization. The β-globin gene is usually used for DNA measurements (e.g., HIV provirus genomes). This requires the use of two primer pairs in a multiplex PCR (Section 29.3.4). One pair amplifies the test DNA, the other a segment of the β-globin gene. Since the amount of the β-globin gene is known and the signal after amplification is constant, this allows conclusions to be drawn about the amount of the target DNA. In contrast to external standards, this procedure allows the detection of inhibitory substances, as long as they have the same influence on both PCRs and are not sequence-specific.

For the quantification of RNA, the signal can be calibrated against the signal from so-called "house-keeping" genes. These are genes that are thought to be expressed in all cells and tissues at the same level at all times. An important aspect that makes the quantitative measurement of RNAs significantly more difficult is the great variation in the efficiency of reverse transcription, in particular when cDNA is synthesized from two different RNAs. A possible way to limit this variance is by the use of artificial standards, so-called mimic fragments.

Competitive (RT) PCR An artificial, cloned standard of known concentration and containing the same primer binding sites is added to the reaction in this procedure. Since it is co-amplified with the same primers as the target sequence and thus "mimics" the target it is called a mimic fragment and the reaction is called competitive (RT) PCR. Ideally the amplified mimic fragment is the same size as the target sequence. After amplification, the two are separated, either by a differential hybridization or by different restriction sites, and analyzed. Alternatively, differentially labeled probes are used to detect the target and mimic products by *TaqMan* PCR (Section 29.2.1).

Competition between the two target sequences occurs whenever the starting amount differs by more than about three to four orders of magnitude (Figure 29.9). The starting sample for measurement is divided into about four equal aliquots and an increasing amount of RNA mimic fragment is added to each. After amplification the two signals are compared to the starting concentration of the RNA mimic fragment and at the intersection of the two curves, the starting amount of the test sample is extrapolated (Figure 29.9).

In general RNA mimic fragments, and not DNA mimic fragments, should always be used, since the greatest variance comes from the RT step, as mentioned above. In addition the RNA mimic fragments should be added at the beginning in order to control each step (sample preparation, amplification, and detection) along the way.

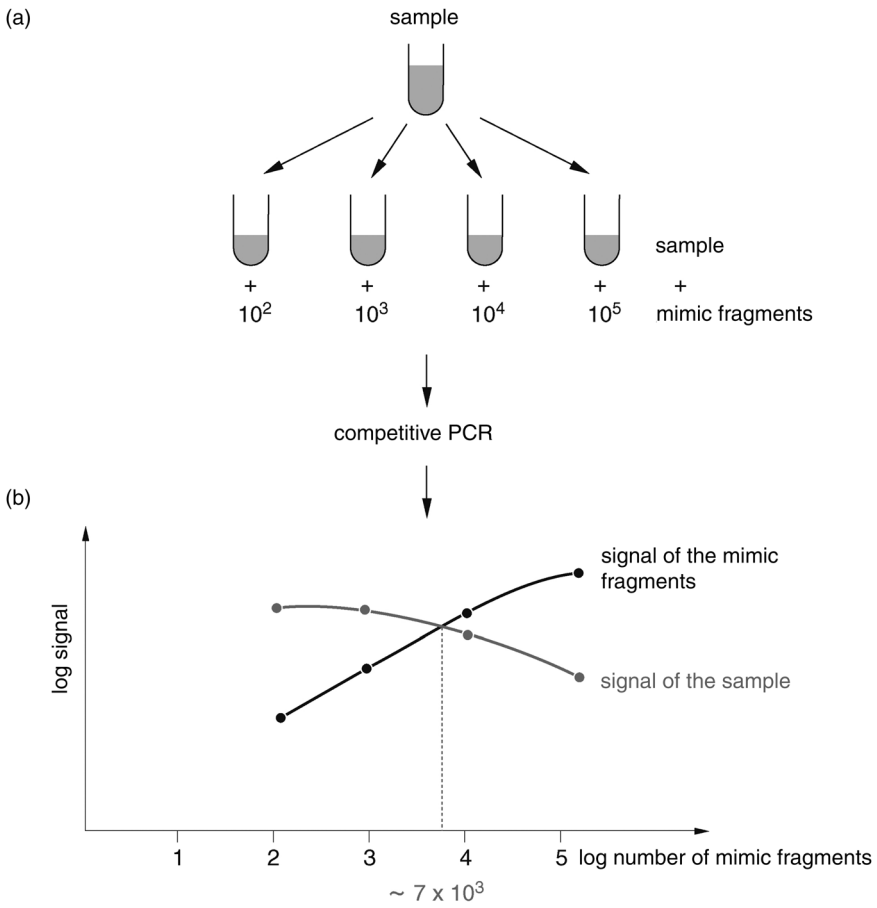


Figure 29.9 Competitive (RT) PCR: The sample to be measured is aliquoted and spiked with an increasing, known amount of mimic fragment. After the amplification, the same pieces of the amplicon are hybridized with the corresponding probe and the signals are compared to the starting concentrations of the mimic fragments. The starting concentration is given by the point of equivalence of the sample.

29.3 Special PCR Techniques

29.3.1 Nested PCR

Nested PCR involves the use of two sets of primer pairs, an outer and an inner pair (Figure 29.10). The advantage of these methods is the increased specificity and sensitivity of the entire reaction, since the outer pair is first used to synthesize a larger amplicon (PCR

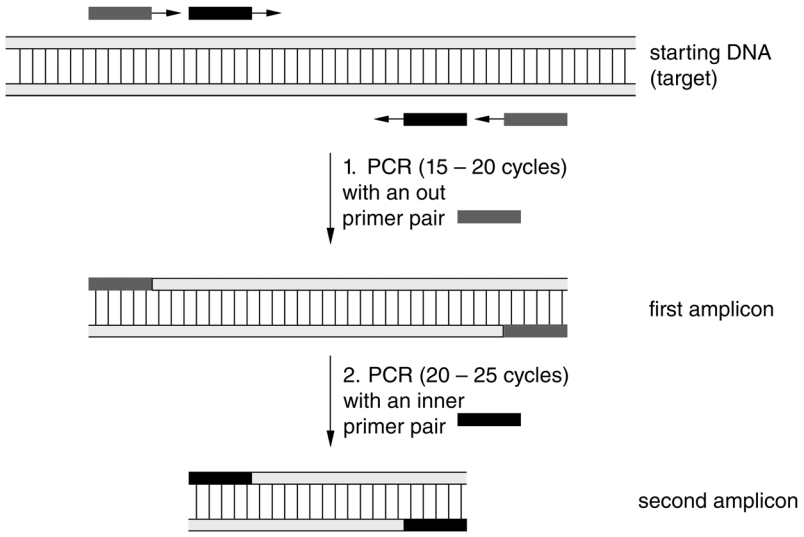


Figure 29.10 Nested PCR: The starting DNA is successively amplified in two separate PCRs. First an out primer pair produces a somewhat larger segment, the first amplicon, then an inner primer pair binds. This pair amplifies a smaller internal segment, the second amplicon, in a further 20–25 cycles. Nested PCR can lead to significantly increased sensitivity and specificity of the PCR.

product) and in a second reaction the inner primer pair is used to amplify the first amplicon, which serves to eliminate the by-products of the first amplification. In general the inner pair is added to the mix after 15–20 cycles and the reaction is allowed to continue for another 20–25 cycles.

The serious disadvantage of this method, however, is the drastically increased chance of contamination (Section 29.4), caused by the pipetting of the amplicon. The possibility of avoiding this danger exists with *one tube nested* PCR, which involves adding all four primers to the reaction at the beginning. The outer primer pairs must have a higher melting point than the inner pair so that at first only the outer pair can anneal under the reaction conditions in use. After the correct number of cycles, the annealing temperature is lowered so that the inner pair can now anneal and produce the inner product.

29.3.2 Asymmetric PCR

When one of the two primers is present in excess relative to the other primer this is referred to as asymmetric PCR. These conditions lead to a selective amplification of one of the two strands. This technique is used for, among other things, sequencing PCR products (Section 29.3.5). If the goal is to hybridize the PCR product with a labeled probe after amplification, it can be advantageous to use asymmetric PCR. The strand that hybridizes to the probe is preferentially amplified, which creates a more favorable situation in the competition between renaturation of the two amplified strands and hybridization with the labeled probe. This comes at the price of the amplification no longer being exponential, but instead it quickly become linear.

29.3.3 Use of Degenerate Primers

Degenerate primers are a mixture of individual molecules that differ at certain points in their sequence. They are used whenever the sequence of the amplification target is not exactly known or the sequences diverge from one another.

The first case comes into play when, for example, a gene segment from a different species, whose sequence is only known for other species, needs to be amplified. Homology searches allow variable positions to be identified and corresponding degenerate primers synthesized that contain all the expected nucleotide variations. They are then used to attempt to amplify the corresponding segment from the desired species. Another use is when only the amino acid sequence of the protein is known. In such cases the amino acid sequence can be used to narrow down the possibilities of the base sequence enough so that degenerate primers can be created. Degenerate primers are also often used when the targets sequences vary from one another. This can occur in the amplification of different HIV subtypes. Even in regions that are otherwise very strongly conserved, single nucleotides can differ from subtype to subtype. The more degenerate the primers are, the greater the danger of nonspecific amplification. This is a fundamental disadvantage of this approach.

29.3.4 Multiplex PCR

Multiplex PCR refers to the procedure of using multiple specific primer pairs to generate the same number of amplicons, all in one tube at the same time. It is apparent that the throughput increases drastically and the amount of work decreases at the same time. Classic uses for multiplex PCR are especially routine diagnostics. For example, the disease cystic fibrosis (CF) is due to certain mutations in the CFTR gene. However, over 100 mutations of this gene are known today, which can be spread across all 24 exons. Multiplex PCR is able to amplify many of the exons at the same time, in order to then investigate the products for point mutations. The situation is similar for other inherited diseases, such as familial hypercholesterolemia, Duchenne muscular dystrophy, polycystic kidneys, and many more.

Another very attractive indication is the diagnosis of several viral infections (HBV, HCV, HIV) at the same time from a single blood sample, which is of particular interest for blood banks. Similarly to the use of degenerate primers, however, it can also be difficult here to get the high complexity of the total reaction under control, which often leads to nonspecific amplification. The newest multiplex protocols have succeeded in measuring up to five viral parameters (HIV-1M, HIV-10, HIV-2, HBV, HCV) at the same time, without nonspecific by-products.

Another example of a multiplexing application is a new test for sepsis, which uses primers that bind to the spacer region of rRNA of pathogenic bacteria to detect and differentiate between a whole palette of pathogenic organisms. This allows the rapid selection of a suitable antibiotic that would otherwise require days with conventional tests, such as selective cultivation of the bacteria, by which time life threatening complications such as septic shock or multiple organ failure may have arisen. The test detects Gram negative (e.g., *Klebsiella pneumoniae*), Gram positive pathogenic bacteria (e.g., *Staphylococcus aureus*), and pathogenic fungi (e.g., *Candida albicans*). The high specificity is demonstrated by the lack of nonspecific cross reactions with over 50 closely related bacteria.

Multiplex PCR is also used for DNA analysis in forensics and paternity testing. Up to 16 PCR amplicons are generated at the same time, which can be separated into different peaks on the basis of different primer fluorescence markers and length after gel electrophoresis. The individual heterogeneity of the amplified repetitive sequence segments show PCR multiplex patterns specific to the individual. These patterns can be stored in binary data banks and rapidly identified with the aid of search programs.

29.3.5 Cycle sequencing

To sequence a PCR product, the sequence need not necessarily first be cloned into phages (M13) or plasmids; instead it can be analyzed directly. Either the product is sequenced subsequent to the PCR or sequencing takes place during the amplification reaction. The latter is referred to as cycle sequencing. Since only one primer is used in each reaction tube, the amplification is linear instead of exponential. As in Sanger sequencing, the chain termination method is usually used, with the aid of dideoxynucleotides. The sequencing takes place, therefore, in four reaction tubes, which differ in the corresponding termination mix (ddATP, ddCTP, ddGTP, DDTTP).

Sequencing According to Sanger: the Did, eoxy Method Section 30.1.1

The reaction can be started with very small amounts of DNA, which is a decisive advantage of the cycle sequencing method compared to older sequencing methods.

In addition, any sort of double- or single-strand DNA can be used as the template. Cycle sequencing is used particularly often for mutation analysis, since it allows the easy investigation of certain genome segments without the need to first clone the region. A disadvantage of this method, however, is that a polymerization error of the *Taq* polymerase that occurs in an early cycle of the linear amplification can be interpreted as a suspected "mutation" and lead to false conclusions. In such cases the opposite strand should always be sequenced. After the reaction is complete, the products are electrophoretically separated on the basis of their differing lengths and the sequence determined. Depending on the label used for the primer, the reaction can be radioactive or non-radioactive. Modern approaches to cycle sequencing utilize differentially fluorescently labeled ddNTPs so that only a single tube reaction needs to be carried out.

29.3.6 *In Vitro* Mutagenesis

PCR is ideally suited to introduce mutations into DNA strands *in vitro* and produce sufficient amounts of mutations to be useful for many purposes. The creation of mimic fragments for competitive PCR (Section 29.2.5) with such a substitution mutagenesis involves the exchange of nucleotide sequences. Mutagenic PCR can also be used to generate a diversity library that can be screened for desired features.

29.3.7 Homogeneous PCR Detection Procedures

To avoid contamination (Section 29.4), real time procedures are increasingly used in closed systems, in which amplification and detection take place in a single step, without the need to

open the reaction vessel. This avoids contamination during pipetting steps or by the formation of aerosols by opening of the reaction vessels.

The use of closed systems and direct detection of the fluorescent signal through the thin glass wall of the capillary tubes or the tops of film-sealed microwell plates at any time during the amplification reaction has greatly reduced the danger of contamination. The formation of amplicons is followed in real time. Another advantage of this homogenous format is its high dynamic range of up to 8–10 orders of magnitude in comparison to heterogeneous formats (see Chapter 28).

Hybridization Methods, Section 28.1.3

29.3.8 Quantitative Amplification Procedures

Besides the quantitative *TaqMan* format (5′ nuclease assay) with *TaqMan*, HybProbe, or molecular beacon probes, homogenous amplifications are also carried out with the aid of fluorescent intercalating dyes. Different amplicons can be detected in multiplex procedures by creating amplicons of different lengths, and thus different melting points, by measuring the resulting hysteresis curves during the PCR cycles.

Another quantitative homogeneous amplification technique is the measurement of amplification products by fluorescence depolarization. In this format, detection takes place through an increase in the polarization, which is a result of binding of the single-strand probes to the generated amplicon. The hybridization of the detection probe to the generated amplicon increases the molecular weight of the resulting complex and thereby the polarization of monochromatic light (e.g., xenon light with a monochromator at 495 nm) shone through it. This technique is also used in association with strand displacement amplification (Section 29.6.2).

29.3.9 *In Situ* PCR

Recently protocols for *in situ* PCR, amplification within cells (e.g., on histological slices) have been published. The difficulty in this procedure is to stabilize the tissue structure on the slide by suitable fixation such that the structure remains intact through the thermal cycles. This is accomplished with special protocols that fix the paraffin-embedded tissue structures with a 10%, buffered formalin solution. *In situ* PCR leads to a large increase in the detected signals in histological slices. There are already thermocyclers on the market that directly heat the slides.

29.3.10 Other Approaches

Besides the special PCR approaches described above, there are many other techniques that are often implemented to answer specific questions or are used for preparative purposes. Only a brief overview can be presented here:

- *Digital PCR:* An alternative approach to quantitate starting copy DNA is to perform limiting dilutions of the sample, followed by hundreds or even thousands of PCRs, such that each reaction contains at most one template copy. Poisson statistics are applied to the fraction of positive reactions to extrapolate the starting concentration without the need for a standard curve. Sophisticated microfabricated devices partition the reaction automatically either into droplets or chambers on a chip, coupled with sophisticated post-PCR detection formats. Such dilution of sample also enables the differentiation and detection of rare mutant sequences.
- *RACE PCR:* rapid amplification of DNA ends is a method to amplify and clone the 5′ ends of cDNAs, in particular those of long mRNAs that were not completely synthesized during *in vitro* transcription.
- *Inverse PCR:* An important method for the amplification of unknown DNA sequences. Primers binding in opposite directions to a known section of sequence are synthesized and used to amplify the DNA in both directions. The products are digested with a restriction enzyme and self-ligated to form circular DNA sequences, which can be amplified and sequenced with the starting primers.
- *Vectorette PCR:* this is used frequently for the characterization of unknown DNA segments.

- Alu PCR: Alu elements are short repetitive elements that are more or less evenly distributed throughout the genome of primates. By amplification with primers that bind to the Alu repeats, a characteristic pattern of bands is created that is so distinctive it can be regarded as a genetic fingerprint of an individual.

cot Curves, Basic Principles of Hybridization, Section 28.1

- *cot Curves:*
- *DOP PCR:* Degenerated optimized PCR (see also Section 29.3.3) is used for the analysis of micro-amplifications and deletions in comparative genome hybridization. In this method, the entire genetic material of a test cell (e.g., cancer cell) and a control cell are amplified and different detection markers are incorporated, such as rhodamine/digoxigenin- and fluorescein/biotin-labeled, Chapter 28). After pre-hybridization of the chromosomes of the target cell with cot DNA, to saturate the repetitive sequences, the mix of digoxigenin and biotin labeled amplicons is hybridized with the chromosomes of the target cell. Detection is accomplished by use of the different filters of a fluorescence microscope for fluorescein and rhodamine. Overlapping the signals reveals a fluorescence pattern which marks those spots where the fluorescein or rhodamine signal predominates due to a micro-amplification or deletion.
- *PRINS PCR:* Primed *in situ* PCR is precursor of *in situ* PCR, in which a primer is extended once after *in situ* hybridization to the target DNA in a fixed target cell.

29.4 Contamination Problems

The high analytical sensitivity of PCR represents, for obvious reasons, enormous progress for science and diagnostics. However, the ability to create many millions of molecules from just a few in a very brief span of time creates an enormous danger of false positive results, since each molecule is an optimal template for further amplification. In addition, the danger of contamination is particularly great when laboratories frequently work with the same primers and always amplify the same target with them. Since PCR will become increasingly common in routine laboratory use in coming years, we will examine the problem in more detail here.

There are three basic types of contamination with DNA:

- cross contamination from sample to sample during isolation of the DNA;
- contamination with cloned material that contains the amplification target sequence;
- cross contamination with already amplified DNA.

In general the greatest danger is presented by aerosols. Table 29.2 demonstrates what sort of contamination is probable from aerosols, which are fine drops of liquid suspended in air. Beginning with the assumed size of such aerosols, the volume of such particles are given. For an amplification reaction with a volume of one hundred microliters, the table shows that already a picoliter contains 10 000 amplifiable molecules that could potentially cause contamination. Each of these amplicons is a perfect target for a new amplification.

29.4.1 Avoiding Contamination

Avoiding contamination should have the highest priority in diagnostic as well in research laboratories, even before decontamination. To accomplish this, it is important to be clear about the possible sources of contamination:

Table 29.2 Danger of contamination by aerosols.

Type of contamination	Size	Volume	Amplicon per volume
Splashes		100 µl	10^{12}
		~1 µl	10^{10}
	~100 µm	~1 nl	10^{7}
Aerosol	~10 µm	~1 pl	10^{4}
	~1 µm	~1 fl	10

- aerosol formation by centrifugation, ventilation, uncontrolled opening of the sample and PCR tubes;
- transfer by contaminated pipettes, disposables, reagents, gloves, clothes, hair, and so on;
- splashes while opening tubes or pipetting liquids.

Numerous measures that can help to minimize the risk of contamination can be inferred from these points.

General measures to minimize the risk of contamination:

- aliquot frequently used reagents and samples;
- only use autoclaved reagents, pipette tips, and reaction tubes;
- minimize manual steps to the greatest extent possible;
- avoid pipetting;
- avoid strong drafts;
- clean and decontaminate devices and pipettes from time to time with dilute bleach solutions;
- if possible, avoid the use of nested PCR, since pipetting the samples drastically increases the risk of contamination (Section 29.3.1).

Sample handling:

- open tubes with cotton wool, a cloth, or something similar to avoid contamination of your gloves;
- if the tube has liquid in the lid, centrifuge briefly;
- work as closed as possible, which means only one tube should be open at a time;
- only use pipette filter tips or positive displacement pipettes;
- change gloves frequently;
- pipette slowly and with care;
- open tubes slowly and with care.

Waste disposal:

- inactivate used (contaminated) pipette tips with HCl or bleach (sodium hypochlorite);
- do not dispose of remaining samples and amplicons in the regular trash, dispose of separately after inactivation;
- close PCR tubes before disposal.

Separation of work areas:

- strict division into three areas:
  - Area 1: preparation of the amplification mixes; for this purpose a laminar flow hood can be used;
  - Area 2: sample preparation;
  - Area 3: amplification and detection, must be in a separate room;
- under all circumstances, separate clothing must be worn in areas 1 or 2 and area 3;
- separate hardware (pipettors, tips, etc.) for each area;
- Sample flow should always be one way: Area 1 → Area 2 → Area 3 → autoclave or waste.

In general, samples and amplified material should be handled with the same amount of care as infectious or radioactive material. In addition, a suitable number of negative controls should be taken through each and every step (sample preparation, amplification, detection) in order to detect contamination early on.

29.4.2 Decontamination

Decontamination includes two different types of measures:

- Chemical or physical measures to clean equipment or the laboratory. This includes substances that destroy DNA directly or at least inactivate it such that it can no longer be amplified. Examples are HCl, sodium hypochlorite and peroxide.

- Measures that are integrated into the routine operation of the tests and take place before or after every amplification. These can be further subdivided into physical, chemical, and enzymatic measures.

Physical Measures UV irradiation: bombarding amplicons after PCR with UV light at a wavelength of 254 nm leads to the formation of pyrimidine dimers (T-T, C-T, C-C) within and between the DNA strands of the amplicon. Such inactivated DNA is no longer usable by *Taq* polymerase as a template. This measure has, however, a few disadvantages. There is a correlation between DNA length and the efficiency of the irradiation: The shorter the amplicon, the less effective the UV light is. In addition, the decontamination is less effective for GC-rich templates than for AT-rich templates.

Chemical Measures Isopsoralens are intercalating dyes, which lead to crosslinking of the two strands when irradiated with long wavelength UV light (312–365 nm). This also blocks the polymerase activity.

3′-Terminal ribonucleotides in primers create a base-sensitive position in the amplicon. A subsequent treatment with base hydrolyzes the primer binding sites.

Enzymatic Measures

- Restriction digestion,
- DNase I digestion,
- exonuclease III digestion,
- UNG system.

Digestion with uracil-*N*-glycosylase (UNG) is the most efficient method for the decontamination of previously amplified DNA. This measure relies on the incorporation of dUTP instead of dTTP during amplification. The resulting PCR product contains uracil bases in both strands and is therefore different than all starting DNAs to be amplified. UNG is an enzyme that cleaves the glycosidic bond between uracil and the sugar phosphate backbone of the DNA. Through subsequent heating or base treatment such abasic DNA hydrolyzes into small fragments and thus can no longer be amplified.

The UNG system is particularly effective for two reasons: first, every newly amplified molecule contains uracil bases and, thus, is a substrate for UNG. Second, UNG decontaminates *before* a new amplification, when possible contaminations are smallest. Many other decontamination measures have the disadvantage that they either occur after the amplification and then need to be quantitatively effective or they require additional steps, which bring with it an additional risk of contamination. Since uracil bases are only substrates for UNG in single- or double-stranded DNA, but not as individual nucleotides or in RNA, the enzyme can be added directly to the amplification mixture and is suitable for use in RT-PCR.

29.5 Applications

Many applications of PCR have already been mentioned in Sections 29.2 and 29.3. Most of these address particular questions in a research laboratory. In the following section, applications in medical diagnostic laboratories will be described in more detail and the possibilities of PCR in genomic analysis will be sketched out.

29.5.1 Detection of Infectious Diseases

The detection of disease causing agents is an ideal application for PCR, since many bacteria and viruses can either not be cultivated at all or only very slowly and conventional tests are not nearly as sensitive as PCR (Figure 29.11). Consequently, such tests have entered into the routine of molecular laboratories in food quality control, as well as veterinary and human medicine. Examples are the viruses HCV (hepatitis C virus), HIV (human immunodeficiency virus), HBV (hepatitis B virus), and CMV (cytomegalovirus), as well as the bacteria *Chlamydia*, *Mycobacterium*, *Neisseria,* and *Salmonella*. By the detection of such pathogenic organisms with PCR, there are three critical aspects: A sufficient specificity of the reaction to avoid false positive results, a very high but also clinically

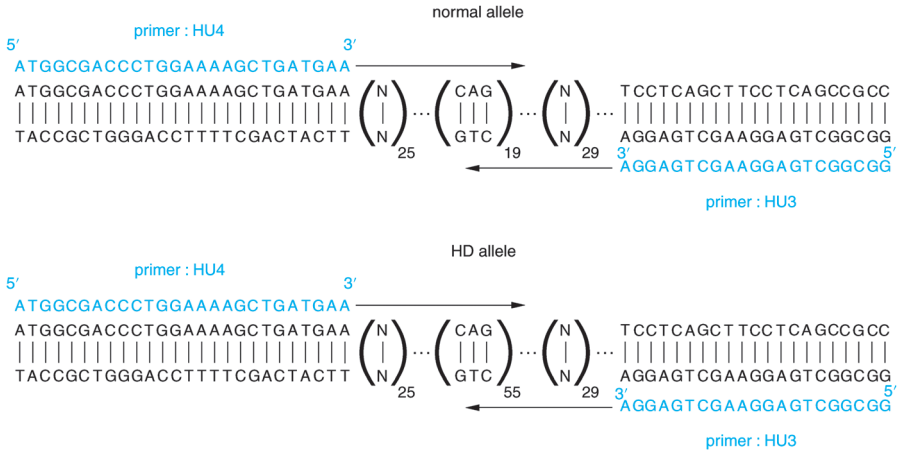


Figure 29.11 Schematic of the detection of the trinucleotide (CAG-) expansion typical for Huntington's disease. The amplification takes place with specific primers that flank the CAG-repeat. The size of the repeats in a polyacrylamide gel provides the diagnostic result (see Figure 29.12).

relevant sensitivity of the test, and a clear, verified result. The challenge of the specificity of the complete reaction results from the question of how specifically a primer pair needs to bind, in order to only amplify HIV but at the same time to recognize all the subtypes, for example. The sensitivity is decisively influenced by the method of sample collection and the volume of the sample. The former must guarantee an efficient separation from inhibitors, to enable an undisturbed reaction. This is particularly important for difficult sample material like sputum, stool, and urine. In addition, the amount of sample naturally has an effect on the sensitivity of the reaction. With ultrasensitive tests such as for the diagnosis of HIV it is often necessary to enrich the virus in the samples. This usually involves ultracentrifugation. In this way a sensitivity of around 20 genome equivalents per milliliter can be achieved. On the other hand, in considering sensitivity, it is always important to factor in the clinical relevance. For example, when a dose of more than 10^5 bacteria from the *Salmonella* group are necessary to trigger an acute gastroenteritis, there is no need for an ultrasensitive test.

The target sequence also plays a decisive role in the accuracy of the test. HIV, like all retroviruses, replicates its genome via a DNA intermediate, the proviruses in the host genome. Only during an acute infection can replicating RNA be found in the blood of the host organism.

Besides the relatively qualitative yes–no answer of PCR, for certain parameters, such as HCV and HIV, more and more quantitative tests are gaining prominence. These allow monitoring of the success of a therapy and thereby help to recognize early on the success of certain therapeutic measures on the course of disease.

29.5.2 Detection of Genetic Defects

In the area of molecular medicine, PCR has provided the prerequisite to diagnose many genetic or acquired diseases on the level of DNA or RNA, prior to the appearance of symptoms. Many methods for this purpose have been developed and refined. In general this is a very new and innovative field that is subject to rapid changes. This section will only provide a general overview of current methods and highlight a few examples.

The detection of known genetic defects can be classified based on the type of mutation, with the exception of translocations, into point mutations or length variations, such as insertions, deletions, and expansions. It is also important to differentiate between simple single site mutations or diseases that are caused by complex mutation patterns. For example, over 300 mutations are known for cystic fibrosis and familial hypercholesterolemia, while Huntington's disease (see below) is caused by a single mutation. Each of these mutation types requires a different method.

Length Variation Mutations Mutated and wild-type alleles of this mutation type can be differentiated on the basis of the length of the PCR product. A well-known example is the detection of trinucleotide expansions for a few neurodegenerative diseases. The causal mutation in Huntington's disease (HD) is the expansion of a trinucleotide repeat (CAG) in the affected HD allele in the IT15 gene. The normal allele already varies in length, with up to 32 repeats being common. Research has shown, however, that a repeat length above 36 CAGs can be described as a positive result. The principle of the test is shown in Figure 29.12. Since the

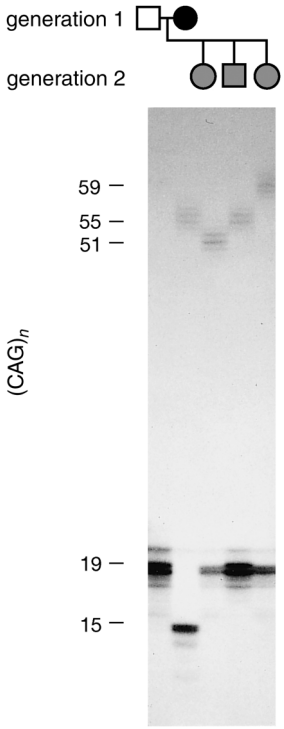


Figure 29.12 Detection of the expansion of the Huntington gene locus. Shown is the trinucleotide expansion ($n = 55$) of an affected woman and her pre-symptomatic children. While her son's expansion is also 55 CAG repeats, transmission of the allele to the first daughter reduced the repeat length ($n = 51$), while the second daughter experienced an expansion ($n = 59$). The healthy father is homozygous, with each allele 19 CAG repeats in length.

disease behaves in an autosomal dominant manner and homozygous carriers practically do not exist, a second, healthy gene is always found. After amplification of both alleles, the PCR products are separated on an electrophoretic gel and the corresponding repeat length is determined (Figure 29.12). Larger length variations, such as those found in Fragile X Syndrome, can also be detected in Southern blots after hybridization with specific probes.

Point Mutations

Sequencing, Chapter 30

Sequencing The surest way to identify and characterize known, as well as unknown, mutations is to sequence the PCR product (Section 29.3.5). Since this is technically demanding and labor-intensive, it is not suitable for screening procedures.

In vitro Restriction and Application, Section 27.1.4

Restriction Fragment Length Polymorphisms RFLPs can be used for analysis whenever mutations have led to the creation or loss of a restriction site. After amplification, the PCR product is cleaved with the corresponding restriction enzyme and the fragments are separated by gel electrophoresis. This method is only suitable for the detection of known mutations.

Reverse Dot Blot (Allele-Specific Hybridization) Reverse dot blots involve immobilizing allele specific probes to the surface of a membrane and hybridizing the PCR product to it. Only in the case of a perfect match between the probe and the amplicon is hybridization allowed. This means that mutated alleles are not bound by the wild-type probe and the wild type does not hybridize to the mutant-specific probe. The location of the hybridization is visualized with specific labels and reveals the gene type. The principle is shown in Figure 29.13. To avoid unspecific binding, the stringency of the hybridization needs to be precisely adjusted (salt concentration, time, temperature). Exact knowledge of the mutations is also required to apply this method.

Specificity of the Hybridization and Stringency; Section 28.1.2

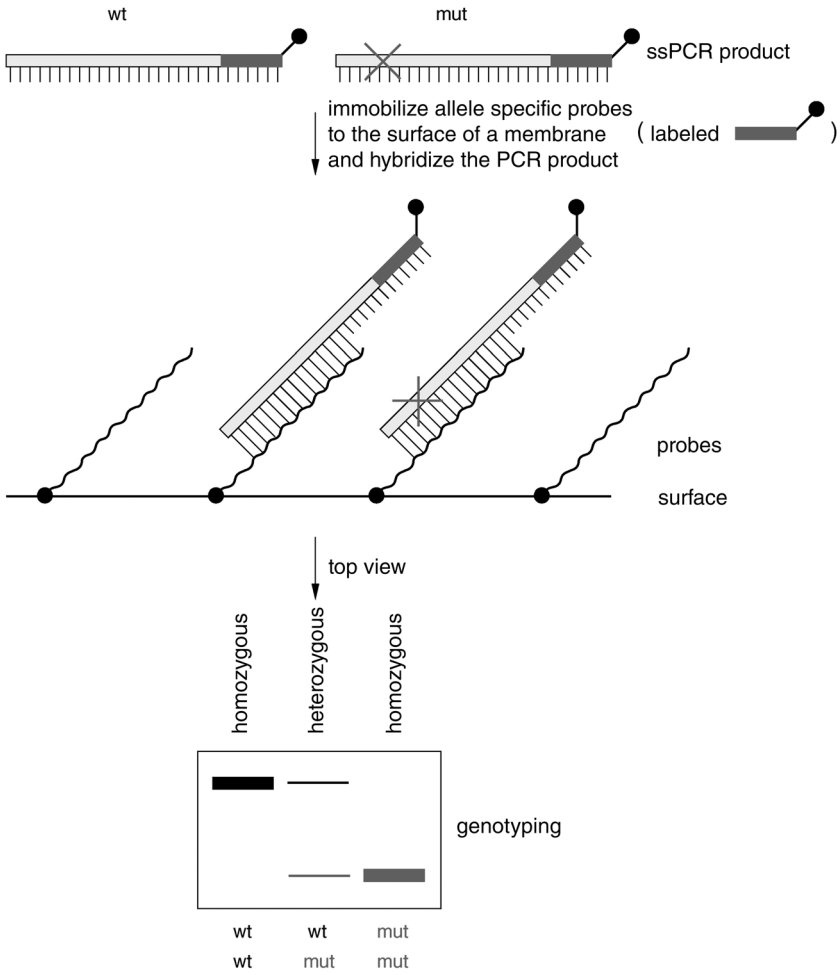


Figure 29.13 Reverse dot blot: The detection of known mutations makes use of allele-specific hybridization probes (wt = wild type; mut = mutant), which are immobilized on a membrane. The position of hybridization is visualized by the label and allows genotyping.

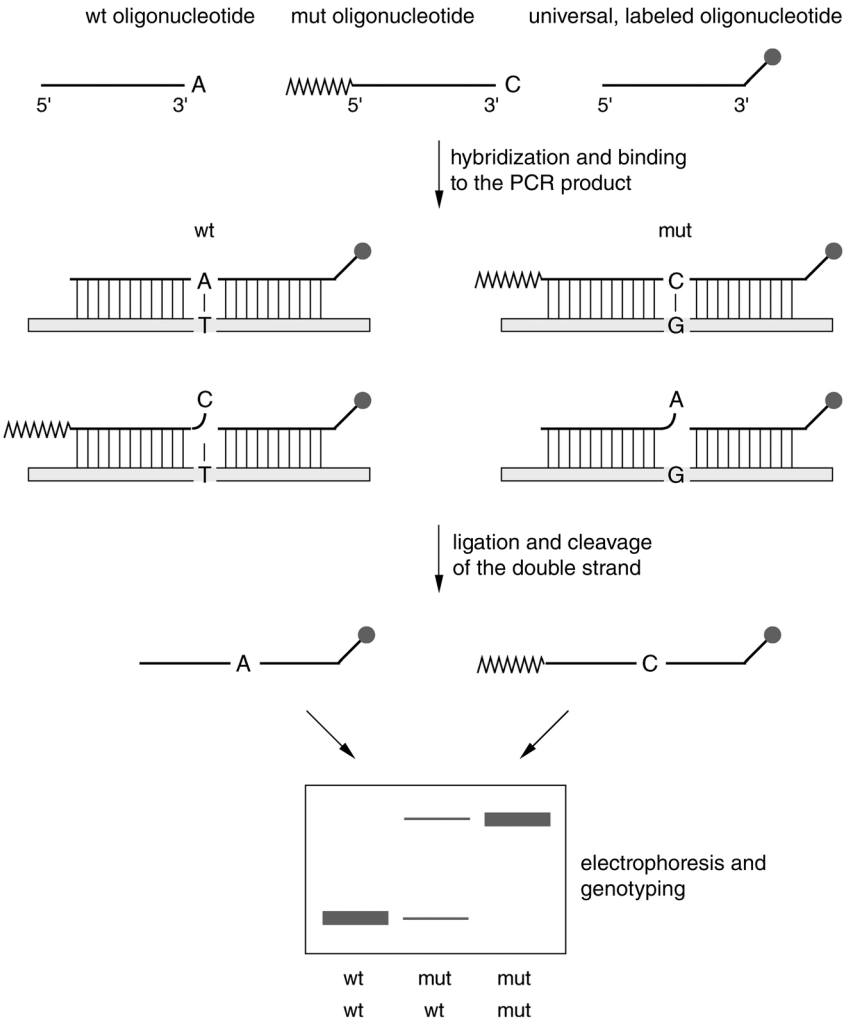


Figure 29.14 OLA (Oligonucleotide Ligation Assay) technique: After a PCR, allele-specific oligonucleotides (wt, mut) bind to the single-stranded amplicon. Only when the 3′ ends hybridize perfectly can the oligonucleotide be ligated to another universal, labeled oligonucleotide. Since the mut oligonucleotide and the wt oligonucleotide differ in length, the ligated oligonucleotides can be separated electrophoretically and detected by their label.

Allele-Specific PCR If the 3′ end of a primer cannot bind to a template, the amplification is inhibited, since *Taq* polymerase only extends a hybridized 3′-OH end efficiently. Allele-specific PCR takes advantage of this fact and uses two different forward or reverse primers. The amplification of the DNA to be investigated takes place in two separate PCR tubes with one of the two sets of primers in each. This allows the elegant characterization of the genotype (homologous wild type, heterozygous, homologous mutants). The exact mutations must be known for design of the primers.

OLA Technique The oligonucleotide-ligation assay (OLA) also takes advantage of the fact that only perfectly hybridized, neighboring oligonucleotides can be ligated together. To analyze a known mutation, oligonucleotides are created that differ in length or labeling, and bind to the PCR product in an allele-specific manner. Depending on the presence of the mutation, the ligase connects one or the other allele-specific oligonucleotides with the universal oligonucleotide. This provides the information for the genotyping (Figure 29.14).

Single-Strand Conformational Polymorphism (SSCP) Single-stranded DNA (ssDNA) forms unpredictable intramolecular secondary structure under renaturing conditions, which is strictly dependent on the primary sequence. Since a mutated allele carries a different sequence, it will also adopt a different conformation upon renaturation. In single-strand conformational polymorphism (SSCP) analysis the PCR product is denatured after amplification and immediately loaded onto a renaturing gel. Even the smallest changes in conformation of the single strands leads to a different mobility in the gel, which is observed as a band shift (Figure 29.15). With this method, unknown

Gel Electrophoresis of DNA, Section 27.2.1

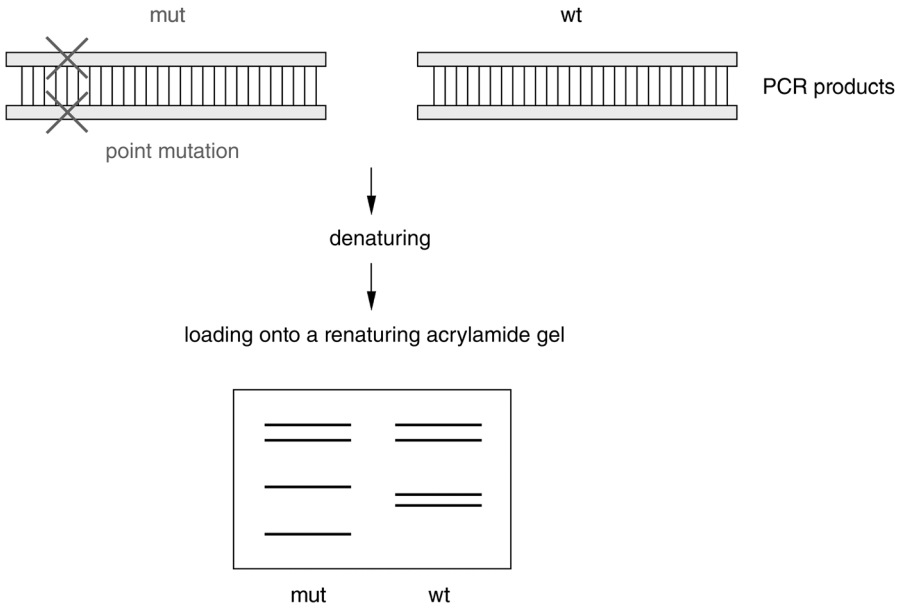


Figure 29.15 SSCP analysis: After amplification the PCR products are denatured and loaded onto a renaturing sequencing gel. Owing to the different refolding, the mobility of the mutated allele (mut) is different from that of the control DNA (wt). In general four bands appear, since the single strand of each allele refolds differently based on its sequence.

mutations or polymorphisms can be recognized, but not characterized. That requires subsequent sequencing.

Denaturing Gradient Gel Electrophoresis (DGGE) This approach is based on a very similar principle to that of SSCP. The double-stranded amplicon is loaded onto a gel that contains a gradient that is more and more denaturing. Depending on the sequence of the allele, denaturation of the mutated spot takes place before or after wild type. The altered mobility is also observed here as a band shift.

29.5.3 The Human Genome Project

In October 2004 the sequence of the human genome was published in the journal *Nature*: The result of 13 years of work involving more than 2800 scientists. An analysis of the data and the 2.85 billion base pairs revealed the presence of 20 000 to 25 000 genes. The standard of quality applied required 99% of the gene-containing sequences to be included and the accuracy was given as 99.999%.

Physical and Genetic Mapping of Genomes, Chapter 36
Generation of a Physical Map, Section 36.2.3

Sequence identification was only the first step, now understanding the function of the genes is the focus.

PCR also strongly drove the development of the Human Genome Project. PCR allowed the introduction and use of sequence tagged sites (STSs), which were a tremendous aid to the mapping work. STSs are specific DNA segments on chromosomes, which are defined by the sequence of two corresponding primers. Such information is readily available through databanks and directly available for use by every researchers involved in mapping the human genome or cloning genes. If an STS is part of an expressed sequence, it is referred to as an expressed sequence-tagged (EST) site. A particular form of STS are short tandem repeat polymorphisms (STRPs), short dinucleotide repeats, usually CA, which can vary in length from individual to individual. STRPs allow the determination of recombination frequencies and therefore conclusions about the separation of such markers. They also make the characterization of haplotypes possible and the isolation of genes by a positional cloning approach. The first gene that was cloned using a positional cloning approach was the CFTR gene, which is responsible for cystic fibrosis, in 1989.

Identification and Isolation of Genes, Section 36.2.4

The modern DNA/RNA sequencing methods (next generation-, deep-, sequencing) would not have been possible without the developments in the field of PCR. The parallelization and miniaturization of PCR were decisive. The use of ever smaller volumes (down to picoliters), allows use of less sample and fewer reagents. Speed and precision also increase. The current focus of development is the commercialization of microfluidics, in which preparative amounts

of nucleic acids can be created for sequencing purposes. By suitable choice of primers, tags and barcodes can be incorporated into PCR amplification. They can then, depending on the sequencing protocol, be further amplified by, for example, emulsions PCR or bridge PCR.

29.6 Alternative Amplification Procedures

Besides PCR, other amplification procedures exist for the multiplication of nucleic acids. Even if such procedures have found a use in individual laboratories in the past, it is becoming increasingly clear that a broad, routine application can only be accomplished through the use of PCR. Therefore, only a few of the most important alternatives will be discussed in this book.

29.6.1 Nucleic Acid Sequence-Based Amplification (NASBA)

This procedure involves an isothermal amplification of nucleic acids at 42 °C. Enzymatic components of the NASBA reaction are reverse transcriptase, RNase H, and T7 RNA polymerase. Another special aspect is the incorporation of a T7 promoter by means of a primer (primer A). This occurs when the promoter is attached to the specific nucleotide sequence of the primer. The reaction is started with the binding of Primer A to the RNA template. Reverse transcriptase (RTase) converts the template into cDNA and the RNA of the hybrid is digested by the RNase H. Primer B then binds to the opposite strand and the DNA-dependent, DNA polymerase activity of the RTase completes synthesis of the strand. This leads to the replication of the promoter. T7 polymerase, as the third enzymatic component of the mixture, binds to the promoter and synthesizes around 100 new RNA molecules (dependent on the length of the amplicon), and the process repeats from the beginning (Figure 29.16). Self-sustained sequence replication (3SR) is based on the same principle. NASBA and 3SR can be slightly modified and DNA templates transcribed into RNA prior to beginning the reaction to allow use of DNA templates.

Transcription-Mediated Amplification (TMA) This procedure is an alternative isothermal amplification of RNA at 42 °C. Enzymatic components of the TMA reaction are a reverse transcriptase and T7 RNA polymerase. RNase H is replaced by the partial RNase H activity of the reverse transcriptase in this reaction. This amplification procedure also involves the incorporation of a T7 promoter attached to a primer (Primer A) into the amplicon.

29.6.2 Strand Displacement Amplification (SDA)

This method of nucleic acid replication also involves an isothermal reaction. It is based on the ability of DNA polymerases to begin new synthesis at single-strand breaks and displace the old strand in the process. Amplification involves a cyclical cleavage of single strands and the subsequent strand replacement. Since restriction enzymes normally completely cleave double-stranded DNA, the creation of the single strand breaks, called nicks, is done by incorporating a nucleotide analog into the opposite strand. After cleavage the single stranded (inactive) remaining sequence of a restriction site is made double-stranded by extending a primer. New synthesis does not employ the four natural dNTPs, which would lead to a double-stranded reaction site subject to complete cleavage, but occurs instead in the presence of three of the natural dNTPs and a thio-deoxy nucleotide. This results in the formation of a double-stranded hybrid between the normal and the newly synthesized sulfur-containing strand. This site cannot be completely cleaved by the restriction enzyme, which leads to the desired single strand nicks. The principle is illustrated in Figure 29.17.

29.6.3 Helicase-Dependent Amplification (HDA)

As an alternative to strand displacement, DNA helicase is employed and the resulting single stranded DNA is protected from reassociation with single-strand DNA binding proteins. In the next step two primers are used, as for PCR, and DNA polymerase generates the two daughter strands. These strands then become available for the helicase and the next round of amplification is begun. In a sense, it is just like PCR at a constant temperature. One advantage is that it

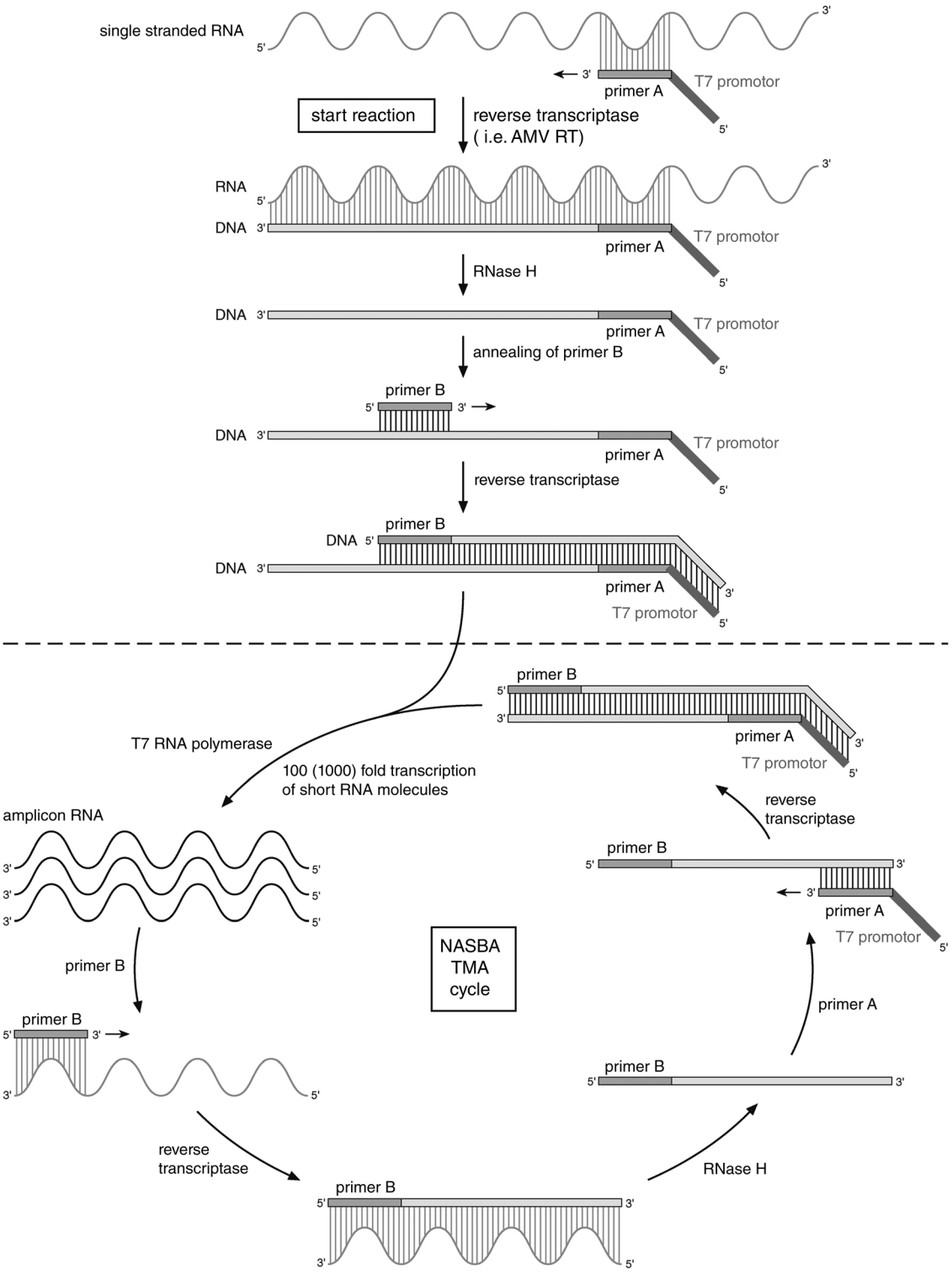


Figure 29.16 Nucleic acid sequence-based amplification (NASBA): The starting point of the amplification is a single-stranded RNA that binds to Primer A. This primer contains a T7 promoter sequence on its 5′ end. A cDNA is synthesized by reverse transcriptase and the RNA in the hybrid is digested immediately by RNase H. Primer B now binds to the single-stranded DNA and synthesizes the opposite strand, in so doing creating a functional T7 promoter. The T7 RNA polymerase recognizes its promoter and synthesizes around 100 RNA molecules, dependent on amplicon length, which then trigger the cyclical phase of the NASBA reaction, in which the described steps are repeated. The complete reaction takes place at a constant temperature and in a single amplification buffer.

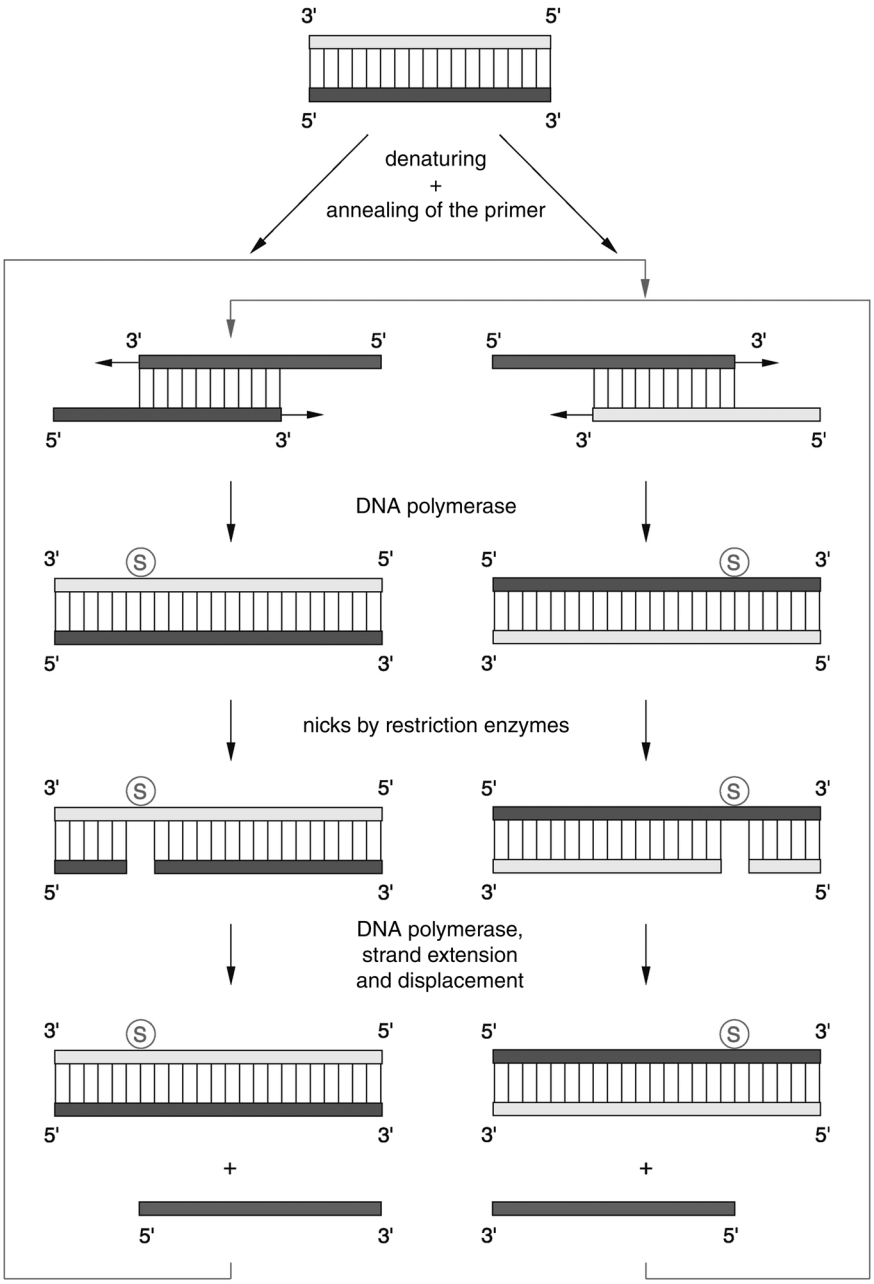


Figure 29.17 Strand displacement amplification (SDA): This is a cyclical process consisting of synthesis, restriction digestion, and strand displacement. The primers contain the recognition sites for the restriction enzyme. Since new synthesis is carried out in the presence of a thio nucleotide, nicks result, since such thio-bonds are resistant to restriction enzymes. SDA, like NASBA, runs under isothermal conditions. Source: according to Persing, D.H. *et al.* (1993) *Diagnostic Molecular Microbiology: Principles and Applications*, American Society for Microbiology, Washington D.C.

does not require a thermocycler. A disadvantage is, however, that the choice of primers and optimization of reaction conditions involves a more complicated search.

29.6.4 Ligase Chain Reaction (LCR)

The ligase chain reaction (LCR) does not lead to increasing the amount of the actual target sequence, instead it leads to an amplification of two oligonucleotides ligated together, which are complementary to the original strands (Figure 29.18). After the initial hybridization of two immediately neighboring oligonucleotides, they are linked together by a thermostable ligase. These form the target for two complementary oligonucleotide pairs, which also hybridize and are linked by the ligase. In about 30 cycles LCR reaches a similar sensitivity to that of PCR. To increase the amplification specificity, LCR protocols have been developed in which the two inner 5′ ends of the oligonucleotides are selectively phosphorylated, to avoid unspecific ligation.

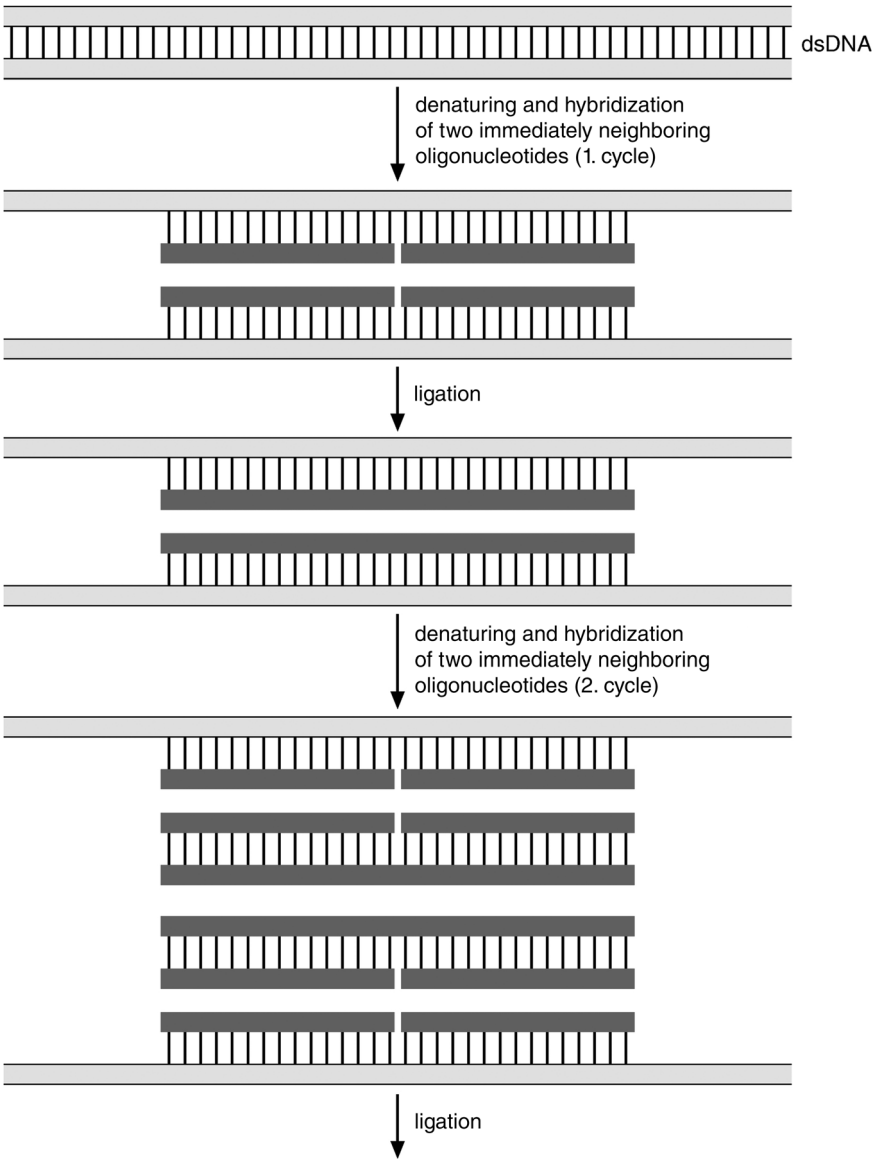


Figure 29.18 Ligase chain reaction (LCR): Theoretically each round of the cycle, consisting of denaturation, annealing of four oligonucleotides, and ligation, leads to the doubling of the number of the oligonucleotides linked together. Analogous to PCR, LCR also leads to exponential amplification.

Repair Chain Reaction (RCR) The repair chain reaction is related to the ligase chain reaction. In contrast to the ligase chain reaction, the two complementary oligonucleotide pairs do not abut, but instead are separated by a gap of one or more nucleotides (Figure 29.19). The gap is selected such that the addition of dGTP and dCTP or dATP and dTTP, a polymerase and the ligase repairs the missing nucleotides of the gap in a double strand-dependent reaction. This combined, limited, elongation and ligation increases the specificity of the amplification, since the two oligonucleotide pairs cannot be ligated together without gap filling taking place.

29.6.5 Qβ Amplification

Qβ amplification does not involve the elongation of a primer (like PCR or NASBA/TMA), instead new synthesis is triggered by the structure of the Qβ genome. Copying the Qβ structures leads to (–) and (+) copies in each reaction cycle and therefore leads to exponential amplification. The proliferation takes place isothermally, after the triggering Qβ are coupled to a parameter-specific probe, which hybridizes with the target sequence (Figure 29.20). A disadvantage of this method is that it is a signal amplification, which can also lead to the amplification of falsely hybridizing probes, causing false positive results.

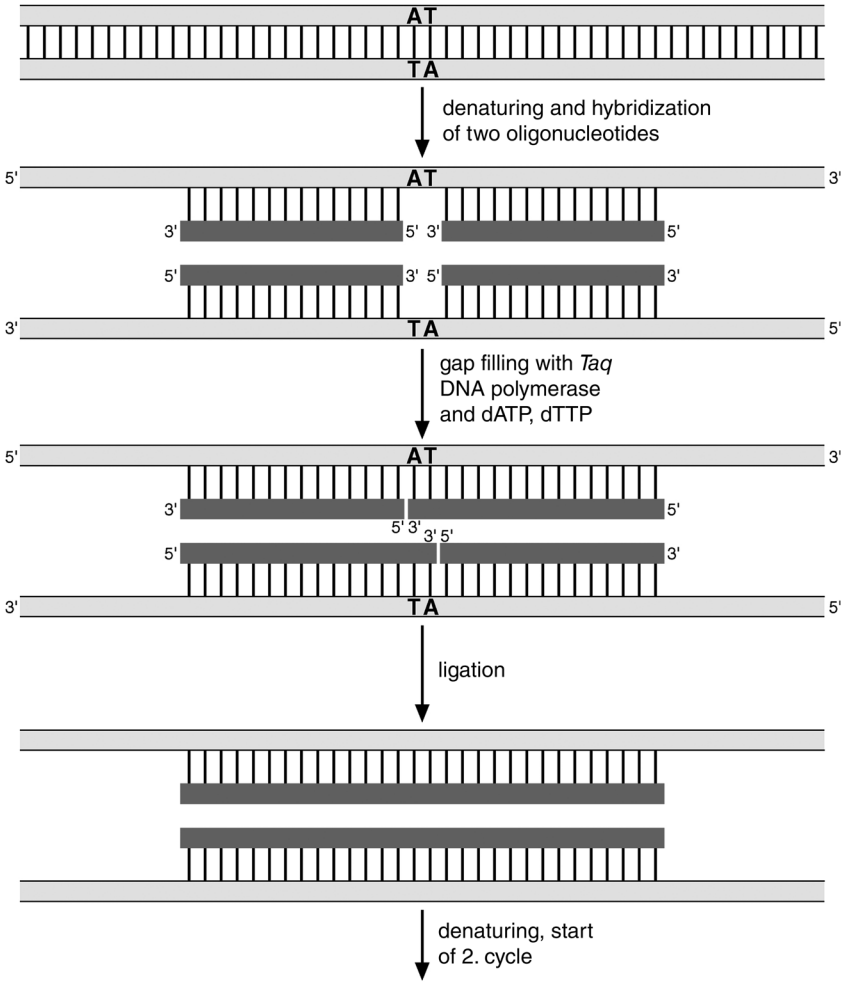


Figure 29.19 Repair chain reaction (RCR): In contrast to LCR, RCR does not lead to the annealing of the two primers next to one another, instead it leaves a gap of several nucleotides. Ligation only becomes possible after a polymerase has filled in the gap in the presence of dGTP and dCTP or dATP and dTTP, which increases the specificity of the reaction. The gap is chosen such that only dG and dC or dT and dA are contained in the opposite strand of the target DNA in the gap.

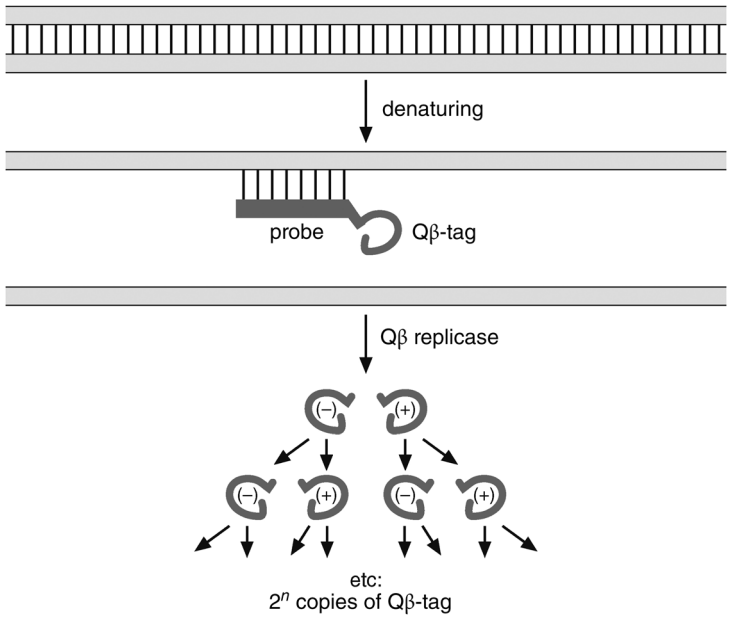


Figure 29.20 Qβ amplification: Qβ amplification involves a structure-initiated replication of the Qβ indicator sequence by the enzyme Qβ replicase at an isothermal temperature. The amplification is exponential and leads to a rapid increase in the number of amplification products through (–) and (+) strand intermediates. In contrast to PCR or other target amplifications, Qβ amplification leads to the proliferation of the Qβ indicator sequence; this is a signal amplification of a coupled, target-independent indicator sequence, not the target sequence itself.

29.6.6 Branched DNA Amplification (bDNA)

Signal Amplification, Section 28.5.3

The branched DNA amplification (bDNA) method is a more recent means by which to amplify signals. The target nucleic acid to be detected is coupled to a solid surface with specific capture probes. Other nucleotides, extenders, then hybridize to the nucleic acids. The extenders bind target-independent amplifier molecules, which then stick out from the original target like antenna. These amplifiers bind many more oligonucleotides, which carry an alkaline phosphatase label. After several washes, the phosphatase substrate is added and chemiluminescence is used to detect the nucleic acids. The bDNA method allows the detection of around 10^5 target molecules. A disadvantage of this method, however, is that it involves a signal amplification, which likewise amplifies the signal resulting from a falsely hybridizing probe and therefore can lead to false positive results.

29.7 Prospects

PCR has become a central bioanalytical method of molecular research laboratories. If current trends continue, it will become increasingly important in routine diagnostics. There are many reasons why PCR is one of the cornerstones of modern molecular biology. Increased automation, not only of PCR, but also of the critical preceding sample preparation, plays an important role. The enzymatic contamination control systems have been particularly important for routine diagnostic use. In addition, closed tube detection formats, such as *TaqMan,* allow routine use with fully automated PCR analysis devices with sensitivity in the range of a few copies. RT-PCR has allowed application to retroviral diagnostics (e.g., for HIV and HCV) and expression profiling.

Another trend is shortening reaction times with faster thermocyclers and minimizing the reaction volume, as has already been realized in the LightCycler®. One also hopes to reduce amplification times down to minutes by miniaturizing the PCR reaction vessel in the form of chips (lab on a chip). The continuous development of microfluidics technology is promising for future routine analytics. Digital PCR offers the possibility of precise absolute quantitation, rare mutation detection, and quantitation of small differences in copy number.

It can be expected that further automation of the entire workflow together with miniaturization will enable different and more challenging sample types, such as whole blood, sputum, urine, and spinal fluid to be amenable to PCR. Future applications that will take advantage of further automation and miniaturization of both sample preparation and PCR are on the horizon. In particular, point of care PCR devices that allow minimally trained operators to perform PCR from sample to answer in what is currently relegated to sophisticated laboratories with highly trained technicians is particularly exciting. Such rapid, simple RT-PCR would facilitate the spread of cost-effective DNA diagnostic methods for diseases such as Ebola and HIV in countries with limited financial resources and the detection of foodborne contaminants and fraudulently labeled foodstuffs on-site.

Further Reading

Dieffenbach, C.W. and Dveksler, G.S. (2003) *PCR Primer. A Laboratory Manual*, 2nd edn, Cold Spring Harbor Laboratory Press, New York.

Logan, J., Edwards, K., and Saunders, N. (eds) (2009) *Real-Time PCR: Current Technology and Applications*, Caister Academic Press.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Kessler, C. (ed.) (2000) *Non-radioactive Analysis of Biomolecules*, Springer, Berlin, Heidelberg.

Larrick, J.W. and Siebert, E.D. (1995) *Reverse Transcriptase PCR*, Ellis Horwood, London.

Yang, S. and Rothman, R.E. (2004) PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings. *Lancet Infect. Dis.*, **4** (6), 337–348.

Rabinow, P. (1996) *Making PCR: A Story of Biotechnology*, University of Chicago Press.

Espy, M.J., Uhl, J.R., Sloan, L.M., *et al.* (2006) Real-time PCR in clinical microbiology: applications for routine laboratory testing. *Washington Clin. Microbiol. Rev.*, **19**, 165–256.

Millar, B.C., Xu, J., and Moore, J.E. (2007) Molecular diagnostics of medically important bacterial infections. *Curr. Issues Mol. Biol.*, **9**, 21–40.
Reischl, U., Wittwer, C., and Cockerill, F. (2002) *Rapid Cycle Real-Time PCR. Methods and Applications*, Springer, Berlin, Heidelberg.
Saunders, N.A. and Lee, M.A. (eds) (2013) *Real-Time PCR: Advanced Technologies and Applications*, Caister Academic Press.
Hugget, J.F., Foy, C.A., *et al*. (2013) The digital MIQE guidelines: minimum information for publication of quantitative digital PCR experiments. *Clin. Chem.*, **59** (6), 892–902.
Niemz, A., Ferguson, T.M., and Boyle, D.S. (2011) Point-of-care nucleic acid testing for infectious diseases. *Trends Biotechnol.*, **29**, 240–250.
UNITAID (2014) *HIV/AIDS Diagnostic Landscape*, 4th edn, World Heath Organization. http://www.unitaid.eu/images/marketdynamics/publications/UNITAID-HIV_Diagnostic_Landscape-4th_edition.pdf.

DNA Sequencing

30

Jürgen Zimmermann and Jonathon Blake

EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany

In 1975, Fred Sanger laid the foundation for the most powerful tool for the analysis of the primary structure of DNA, with the development of an enzymatic sequencing method. At that time, neither the far-reaching implications for the understanding of genes or whole genomes nor the rapid development of this method could have been foreseen. Back then, Fred Sanger was happy about the sequencing of five bases in one week, as he himself noted in retrospect at a reception in the Sanger Center (Cambridge, England) in 1993.

In comparison to these five bases, genome sizes reach astronomical dimensions. The average length of a small viral genome is in the range of 10^5 base pairs (bp). With the increasing complexity of organisms, further magnitudes are quickly exceeded: *Escherichia coli* already reaches 4.7×10^6 bp, *Saccharomyces cerevisiae* 1.4×10^7 bp, *Drosophila melanogaster* 1.8×10^8 bp, and humans 3.2×10^9 bp. Genomes of plants and even lower organisms can achieve even greater lengths: wheat (*Triticum aestivum*) 1.6×10^{13} bp and *Amoeba dubia* 1.2×10^{14} bp. The actual number of bases to be sequenced can easily reach hundred times the genome size depending on the strategy used. In this consideration, the needs of diagnostic DNA sequencing, which are increasingly gaining in importance, are not yet included. In addition, even though DNA sequencing analyzes only small fragments, it processes these in large numbers.

At the same time as the development of the Sanger method, cloning methods became available in M13 phages, which allowed both the biological amplification of DNA fragments in a size range of up to two kilobase pairs as well as the generation of "easily" sequenceable single-stranded DNA. A maximum read length of 200 bp could be achieved. Therefore, only a fraction of the entire sequence could be determined during a sequencing run. This disparity forced the development of sequencing strategies that – at a reasonable expense – enabled the reconstitution of the whole sequence.

Equipped with the tools of the Sanger method, it began with the analysis of whole genomes in the 1970s. In 1977, Sanger and his coworkers published the 5386 bp DNA sequence of the phage phiX174. In 1982, the complete sequence of the human mitochondrion with a length of 16 569 bp was determined. By 1984, they had already achieved, with the sequence of the Epstein–Barr virus, a length of 172 282 bp. Only 25 years after the breakthrough of Sanger, in 2001, the first sequence of the human genome was published. The cost of the Human Genome Project (HGP) exceeded the limit of one billion US dollars by far.

The project made one thing clear: a cost-effective and rapid sequencing of larger genomes would have hardly been possible with existing technologies, and a lot of questions remained unanswered, such as the dynamics of genomes, even if the data constituted a milestone and the results made many procedures possible, such as microarray analysis. It would take many years for massively parallel sequencing (MPS) approximation methods, also known as next generation sequencing (NGS), to become available. The gross output of 300 Gb and more per experiment and the associated – and substantial – cost reduction created the chance for new procedures to allow for new objectives covering all aspects of genomes, which were previously not possible (Table 30.1).

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

Table 30.1 Application areas in the field next generation sequencing (NGS).

Structural genomics	***De novo* sequencing**	
	Resequencing	SNPs, structural variants, exome, tumor/normal tissue, personalized medicine, GWAS
	Metagenomics	
	Pharmacogenomics	
Functional genomics	**Transcriptome sequencing**	
	sRNA	miRNA
	Protein–DNA interactions	ChIP
	Epigenetic modifications	DNA methylation

The dynamics of the new sequencing systems is found in the arbitrarily increasing coverage/depth, with which the areas are analyzed.

In a microarray experiment, this is limited and it is difficult to find rare transcripts in a sample. With a sequencer, it is in principle easier to detect rare transcripts by manipulating the variables mentioned above.

Most of these methods are still based on the Sanger method in their basic approaches, and despite their substantial improvement in throughput previous sequencing processes are still used in their own right, for example, for sequencing individual genes to validate expression constructs and PCR products. In addition, it should not go unnoticed that the new methods have their own systematic errors: enzymatic amplification is necessary in most cases to obtain enough products that can be detected in an instrument later on. These include preferences in the amplification techniques used, a guarantee that the generation of sequencing libraries did not occur fully randomized (e.g., negative GC-selectivity in the generation of sequencing libraries) or even preferences for sequencing enzymes that distort the results. This may result in a non-representative occurrence of all sequence regions; hence a complete picture is questionable. Even in methods that are not based on amplification the instruments themselves suffer from selectivity. While the MPS method is reaching an increasing methodological maturity, the first single molecule sequencing devices are already available and other methods are in development.

In 1996, more than 300 mega-base (Mbp) of sequence information was newly recorded in the EMBL Nucleotide Sequence Database. This is nearly as much as had been registered in the 13 years since the founding of the institution. In June 2005, 95 giga-bases (Gb) in 54×10^6 entries could be published. This is the same as the daily production of an MPS machine. As of March 2011, there are 206×10^6 entries and 319×10^6 bp.

These figures illustrate both the technological advances of the processes used as well as the increasing use of these techniques. As great as the numbers may seem, they represent sequences from different organisms, different versions, and also sequence fragments of small size, of which the location is not known in all cases. The path to a final single and complete sequence as the representation of a genome is not to be underestimated and requires a considerable effort. Recently the first projects for "Platinum genomes" have been started.

30.1 Gel-Supported DNA Sequencing Methods

As already mentioned, gel-supported methods are still broadly used in DNA sequencing of smaller portions and diverse objectives such as the screening of expression constructs, PCR products, or structures that do not parallelize for use on a MPS (massive parallel sequencing) device, have overly low length, or cannot be resolved on the new sequencers at all until now.

The original *shotgun* method was based on the statistical reduction of the genome into small fragments of 1–2 kb in length, their cloning, and sequencing as well as the assembly of the individual sequences like a puzzle. The overall picture of the sequence thus arises only at the end of a project. This uncertainty during the project was addressed by developing orderly and

purposeful methods. *Primer walking*, *nested deletions*, Delta restrictions cloning, *chromosome walking* or combinations of methods allow, even during sequencing, a localization of the obtained information. Hybrid procedures combine the high initial data rate of accumulation of a random strategy with the reduction of the sequencing effort of a directional strategy. The starting point for genome-wide DNA sequencing is a correspondingly precise physical map that acts as a rough guide. Positions and relations of individual clones are reviewed with *fingerprinting* methods at a large scale and with *fine-mapping* they are reviewed in detail.

Contig A continuous (contiguous) DNA sequence that is generated during the mathematical assembly from overlapping DNA fragments in the computer.

The entire process of cloning and sequencing is based on statistical events that lead to an unpredictable sequence representation of individual results. Accordingly, gaps between sequence contigs are to be expected. These may have their origin both in sequence gaps (e.g., a too short read length) as well as in physical gaps (section missing from the corresponding clone library).

Sequence gaps are usually closed by primer walking on the cloning, in which the neighboring but not associated contigs come to rest.

Physical gaps can only be closed with the aid of a second clone library, one that was created using a different cloning vector, size selection, and fragment generation. Using an independent library is usually necessary to bypass possible instability of target sequences in a vector/host system. For identification of the missing sequences, the second library is sampled with oligonucleotides (PCR) whose sequences correspond to the ends of the previously associated contigs. A PCR product can arise only where the two end sequences come to rest. The corresponding clone thus contains the missing sequence to connect the two contigs in question.

Sequence repeats (*repeats, inverted repeats*), which exceed the reading length of a read, can also make it difficult to reconstruct the original sequence. Only the use of additional map information of different resolutions as well as the examination of 5′ and 3′ border sequences can help in such cases.

Primer walking

Primer walking is a directed DNA sequencing strategy. In a first step, the DNA fragment that is to be sequenced will be sequenced from both ends in one reaction. Starting from the obtained information, a new primer is placed in the same reading direction at each end. In this way a prolonged sequence section can be determined in each step. In addition, primers are placed in the opposite reading direction to determine the sequence of the complementary DNA strand. At each point of a primer walking project, the position and the running direction of the sequence is uniquely determined. The redundancy of the obtained sequence reaches a value close to two in the optimal case. The sequence information, however, can only be generated serially and in steps and is dependent on the synthesis of a new primer for each reaction.

Nested deletions

The generation of unidirectional nested deletions was developed by Steven Henikoff in 1984. This sequencing strategy allows an ordered sequence extraction using merely a standard primer. The starting material uses plasmids with a known priming sequence, which must have a certain structure: Between the priming point and the cloned insert, there must be singular recognition sites for two different restriction endonucleases whose sequence is not reflected in the insert itself. The cutting site, which is closer to the direction of the priming point, must generate a 3′ overhang. The second, closer to the direction of the insert, must generate a 5′ overhang. After a double digestion with both restriction endonucleases, the linearized DNA molecule is treated with exonuclease III (Exo III). Exo III attacks the 5′ overhanging end of a DNA double strand, where it acts as 3′-5′ exonuclease and successively removes the 3′ recessive DNA. At specific time points, an aliquot is removed and processed parallel in the next steps. In the next step, the 5′ overhang is removed by S1 nuclease to produce a blunt end. After reparation of ends, recircularization and transformation plasmids occur, whose inserts are shortened to a certain number of bases. The sequence shortening takes place right next to the binding site for the sequence primer. Thus, after the primer, a section of unknown sequence begins. Fragments with a maximum length of about 3 kb can be processed. A second series of deletions must be generated to determine the sequence of the complementary DNA strand.

Delta restrictions cloning

Delta restrictions cloning allows for the generation of numerous sub-clones – whose position to each other is known – through simple digestion with polylinker restriction endonucleases. In the first step, the

Cosmids Circular DNA molecules such as plasmids. The name cosmid refers to the DNA sequences that are referred to as *cos* and are derived from the phage lambda. These sections make it possible to add larger genes to cosmids than to plasmids.

BAC Bacterial artificial chromosome (**BAC**) comes from the single-copy F-plasmid of the bacterium *Escherichia coli* and permits stable cloning of longer inserts of more than 300 kbp in bacteria.

fragment that needs to be sequenced is cut and analyzed on an agarose gel. For all clones, which have an internal restriction site for a polylinker enzyme, a piece is removed by this digestion. Simple recircularization of the fragments results in clones that have a deletion relative to the position of the primer and thus deliver *de novo* sequences during the reaction with standard primers.

Transposon-mediated DNA sequencing

This method enables the introduction of primer binding sites in a simple enzymatic step, which in turn allows the sequencing of longer DNA sections without subcloning or primer walking. As means of support transposons are used, which provide both the primer binding and encode a selectable marker (e.g., kanamycin, tetracycline) and allow for bidirectional sequencing. The insertion is carried out by simple incubation with the corresponding transposase and the DNA target mixture. After transformation into *Escherichia coli,* the selected clones can be selected and sequenced. In principle, the systems have no sequence-specific preference. However, in some cases accumulation in the area of certain hotspots can occur.

Possibilities of PCR, Section 29.1

While the sequencing of whole genomes or genes was carried out with classical sub-cloning strategies as a *shotgun* with a framework of directional clones until the turn of the century, in phylogenetic analyses and in clinical diagnostics direct template production through PCR has prevailed. The use of PCR products as sequencing templates only requires a simplified purification (e.g., silica material or magnetic particles) prior to the actual DNA sequencing reaction.

Parallel to the development of sequencing strategies was the development of sequencing techniques. While the first sequencing reactions were still performed with radioactive markers, current techniques make use of fluorescent markers. In addition, the increased understanding of DNA polymerases led to the use of other and new enzymes. After initial work with DNA polymerase I and the Klenow fragment and the use of genetically modified T7 DNA polymerases, thermostable DNA polymerases or even mixtures are now being used. Meanwhile, read lengths of 1000 bp and more can be achieved.

The classic retrieval of RNA/DNA sequences occurs primarily in six steps:

1. *Isolation and purification of nucleic acid:* Genomic DNA is extracted with an adapted method for the target organism(s) and/or the target region. The direct sequencing of RNA is used rarely nowadays. As a rule, a cDNA copy is generated and subsequently sequenced through reverse transcription.
2. *Cloning or PCR amplification:* The DNA obtained in the first step is in several respects not suitable for DNA sequencing: The length of the genomic DNA is too large to be able to process with a conventional procedure. At present, a sequence of ~1 kb can be produced in one sequencing run but only in favorable cases. During the analyses of, for example, a human gene of 50 kb, one only gets a fraction of the total information. For this fragment, it is still necessary to arrange its position and reconstruct it to the original sequence with the other generated fragments. For all types of DNA isolates, the number of copies contained in a preparation is not sufficient for sequencing. The automated DNA sequencing systems that are currently available have a detection limit of about 10 000 molecules, which cannot be achieved in a simple DNA isolate. To obtain a sufficient output amount for the DNA sequencing, it is subject to amplification. For the reconstitution of longer DNA sequences (>2 kb) this process must be conducted *in vivo* in cloning systems. The existing limitations of this process can be avoided with appropriate sequencing strategies. For shorter sequence segments that are frequently analyzed in medical diagnostics, PCR reactions are sufficient.
3. *DNA purification for the sequencing reaction:* To obtain optimal DNA sequencing results, a further purification is required. Contaminating proteins, carbohydrates, and salts can influence the set environment in an uncontrolled manner and can lead to dramatically reduced read lengths. For a more detailed description of the procedures, please refer to the relevant chapters of this book.
4. *DNA sequencing and electrophoresis:* The reaction products of the sequencing reaction of a DNA fragment are fractionated by gel electrophoresis, the generated band patterns are recorded online or offline and subjected to the following analysis.
5. *Reconstitution of the original sequence information:* As mentioned above, the sequence generated in a sequencing reaction is in most cases smaller than the entire sequence that must

be determined. From many fragments generated by sequence reactions, the original image must be restored in the form of a genetic puzzle. For this, computerized and automated methods are used.

6. *Error correction and sequence data analysis:* The obtained sequence information is subjected to quality control. To suppress possible errors of sequencing from individual experiments, the DNA sequences that are to be determined become multiply redundant and, in addition, both complementary strands are sequenced. Subsequently, the sequence is examined for possible contamination by vector, bacterial, or foreign DNA in a first step and, where appropriate, removed from the sequence. With the aid of codon usage tables, potential ORFs (open reading frames), other reference sequences as well as various other tools, possible sequencing errors can be detected. These steps ensure the correction during the sequencing and the subsequent assembly of errors that have arisen. The final generated sequence can be compared to the DNA sequences of already known databases and can undergo a detailed sequence analysis.

The overall process, as outlined in the list above, requires the combined use of a multiplicity of methods that have been described in detail elsewhere in this book. The following sections of this chapter are limited to the actual processes of gel-supported DNA sequencing methods, as well as their labeling and detection methods.

Gel-supported DNA sequencing methods are mainly based on the production of base-specific terminated DNA populations that are separated according to their size in a subsequent, denaturing polyacrylamide gel/linear acrylamide gel electrophoresis. It is therefore basically an endpoint analysis. Up to 96 samples can be processed parallel and read lengths of up to 1000 bp per sample can be achieved. These populations can be generated in two different ways: The reaction products can be prepared by synthesis of a DNA strand (dideoxy sequencing, Sanger method) or by base-specific fission (chemical fission, Maxam–Gilbert method). Gel electrophoresis is performed in capillaries filled with linearly polymerized acrylamide gels.

30.1.1 Sequencing according to Sanger: The Dideoxy Method

The dideoxy method (also known as the chain termination method, terminator method) is based on the enzyme-catalyzed synthesis of a population of base-specific terminated DNA fragments that can be separated by gel electrophoresis according to their size. From the resulting band patterns of a denaturing polyacrylamide gel, the sequence can be reconstructed. The basic principle will be described in the following section.

Starting from a known start sequence, the synthesis of a complementary DNA strand is initiated by adding a sequencing primer (a short DNA oligonucleotide of about 20 bp), a nucleotide mix, and a DNA polymerase. To detect the reaction products, they must be labeled with either radioactive isotopes or fluorescent reporter groups (Section 30.1.2). On one hand, the use of a primer is necessary to obtain a defined starting point for sequencing, and on the other hand it is necessary for the formation of the initiation complex and thus to initiate the start of synthesis of the DNA polymerase. The reaction is started simultaneously in four parallel aliquots, which differ only by the use of different nucleotide mixes. Reactions labeled A, C, G, and T contain a mixture of the naturally occurring 2′ deoxynucleotides and in each case only one type of synthetic 2′,3′-dideoxynucleotides, the so-called terminator (Figure 30.1).

During strand synthesis through stepwise condensation of nucleoside triphosphates in the 5′-3′ direction, two different reaction events can occur: During the condensation of the 5′-triphosphate group of 2′-deoxynucleotides (dNTP) with the free 3′-hydroxyl end of the DNA strand, an elongated DNA molecule is produced with the release of inorganic diphosphate (pyrophosphate), which in turn has a free 3′-hydroxyl group. The synthesis can therefore be continued in the next step (Figure 30.2a).

If, however, a condensation reaction occurs between the free 3′-hydroxyl end of a DNA strand and the 5′-triphosphate group of a 2′,3′-dideoxynucleotide, an extension of this strand is no longer possible, because a free 3′-hydroxyl group is no longer available. The strand is terminated (Figure 30.2b).

The characteristics of the DNA polymerase used and the structure of dideoxynucleotide determine the mixing ratio, leading to the production of a population of base-specific terminated

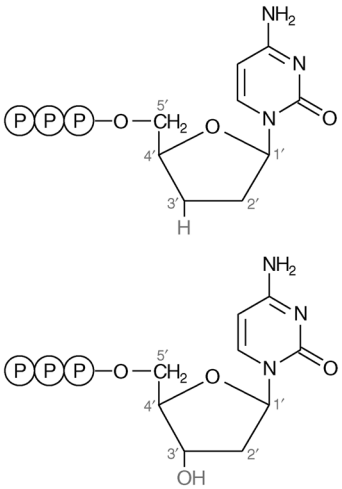


Figure 30.1 Structural comparison of a 2′,3′-di-deoxynucleotide and a 2′-deoxynucleotide. The illustrated 2′,3′-di-deoxy-CTP nucleotide analogue differs from its naturally occurring analogue 2′-deoxy-CTP by the lack of the hydroxy group at C3′ of the sugar.

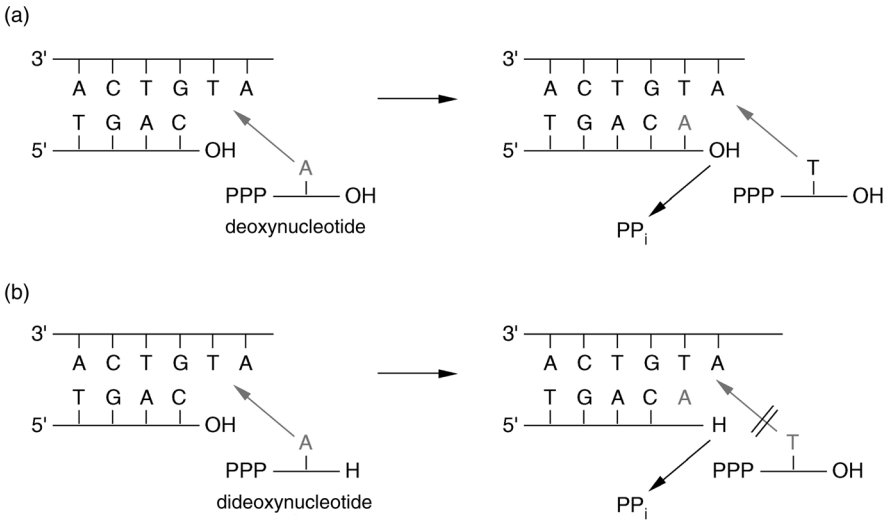


Figure 30.2 Synthesis of a DNA strand with incorporation of a 2′-deoxynucleotide (a) and a 2′,3′-dideoxynucleotide (b). In case of the latter, further polymerization is no longer possible.

reaction products that in each case differ only in one base. At a molar ratio of dNTP : ddNTP of 200 : 1 termination events are relatively rare in the case of catalyzed T7 DNA polymerase reactions. Long reaction products of up to 1000 bp in length are then created. As mentioned above, the reaction is carried out in four aliquots. Each of these partial reactions contains only one type of terminator (ddATP, ddCTP, or ddGTP, ddTTP), which statistically replaces each one of the naturally occurring nucleotides in the synthesized DNA chain. In every reaction vessel products are generated that only end on one base type, such as A.

Gel Electrophoresis of DNA, Section 27.2.1

The reaction products are separated by electrophoresis according to their size in a denaturing polyacrylamide. Gels of the labeled reaction products in all four partial reactions create a "ladder" of bands that each differ by one base (level). From this series of rungs of reactions A, C, G, and T, the base sequence can be read from the bottom (position 1) upwards (Figure 30.3).

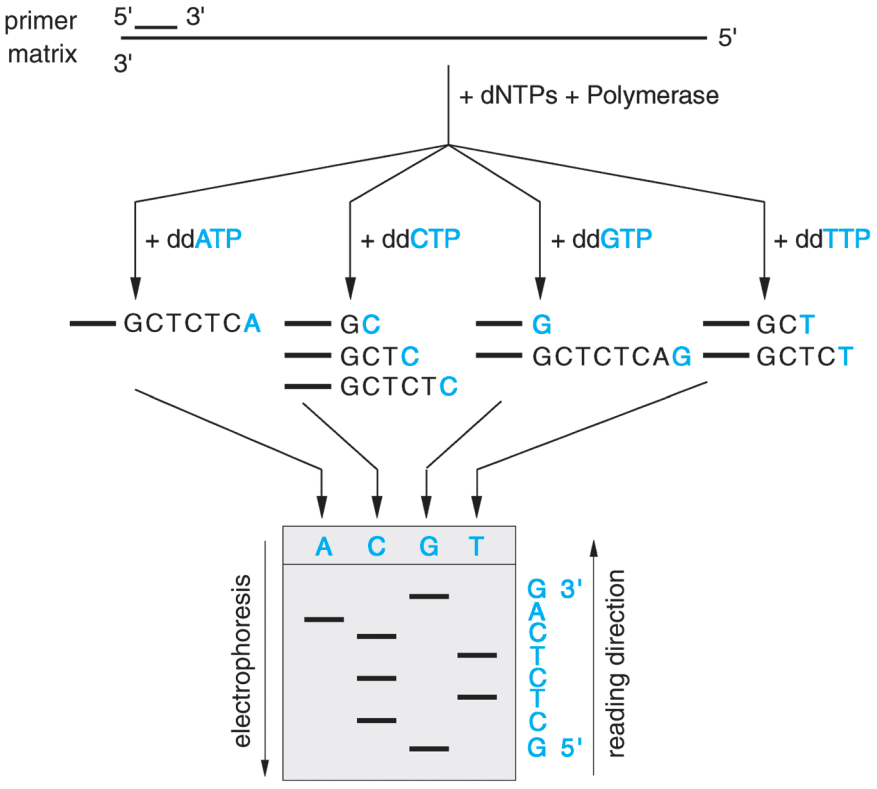


Figure 30.3 Principle of the chain termination method according to Sanger. In a primed DNA synthesis reaction catalyzed by a polymerase, base-specific terminated DNA fragments of different lengths are synthesized. These fragments produce a particular band pattern in the gel electrophoresis that is used for reformation of the base sequence.

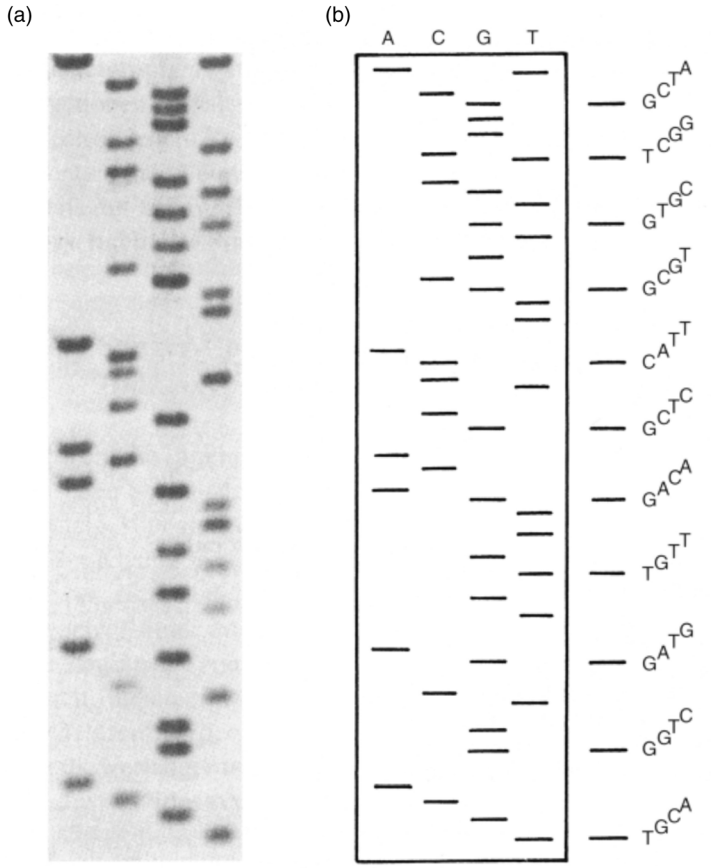


Figure 30.4 Autoradiogram of a sequencing run. In each case four adjacent tracks represent the partial reactions A, C, G, and T of a sample. According to the running direction of the gel, the smaller reaction products are at the lower end of the figure. The sequence is read from bottom to top. Each band is different, ideally by one base from the previous one. Source: adapted according to Nicholl, D.S.T. (1994) *An Introduction to Genetic Engineering*, Cambridge University Press.

Figure 30.4 shows a classic, old-fashioned but instructive, autoradiography of a sequencing gel with radioactively labeled reaction products. The other figures in this section show pseudo chromatogram and fluorescently labeled reaction products. In this illustration, the bands of a track are connected through a cut line in the running direction of the gel and the corresponding band intensities are determined. In this way, the data is reduced by one dimension, creating a representation known as trace data, which illustrates the intensity curve according to time.

This reaction principle has remained unchanged since the developments of Sanger from 1977 until today. The discovery and modification of DNA polymerases led to the refinement of the method described above to protocols in which T7 DNA polymerase is used, and to the development of cyclical process that combine signal amplification and sequencing into one reaction.

T7 DNA Polymerase-Catalyzed Sequencing Reactions T7 DNA polymerase belongs to the class I DNA polymerases, such *as Escherichia coli* DNA polymerase I and *Taq* DNA polymerase. Nevertheless, they differ in their function, their properties, and in their structure. T7 DNA polymerase is the replicating enzyme of T7 phage genome, while the other polymerases are responsible for repair and recombination. Accordingly, the native enzymes contain exonuclease activity: T7 DNA polymerase has a 3′-5′ exonuclease activity, DNA polymerase I 3′-5′ and 5′-3′ exonuclease activity, while *Taq* DNA polymerase shows only a 5′-3′ exonuclease activity. Through N-terminal deletion, the 3′-5′ exonuclease activity could be deleted in T7 DNA polymerases and some thermostable DNA polymerases, as it would otherwise lead to deterioration in the sequencing results. T7 DNA polymerase is distinguished from other non-modified enzymes by their significantly lower discrimination against modified nucleotides. This has the advantage of lower costs and improved sequencing results. It is the ideal model DNA polymerase for DNA sequencing, the stability at higher temperatures is low, however.

T7 DNA polymerase, because of its higher and better distribution of signal intensity of sequence patterns and the low signal background, has made the Klenow fragment, which was

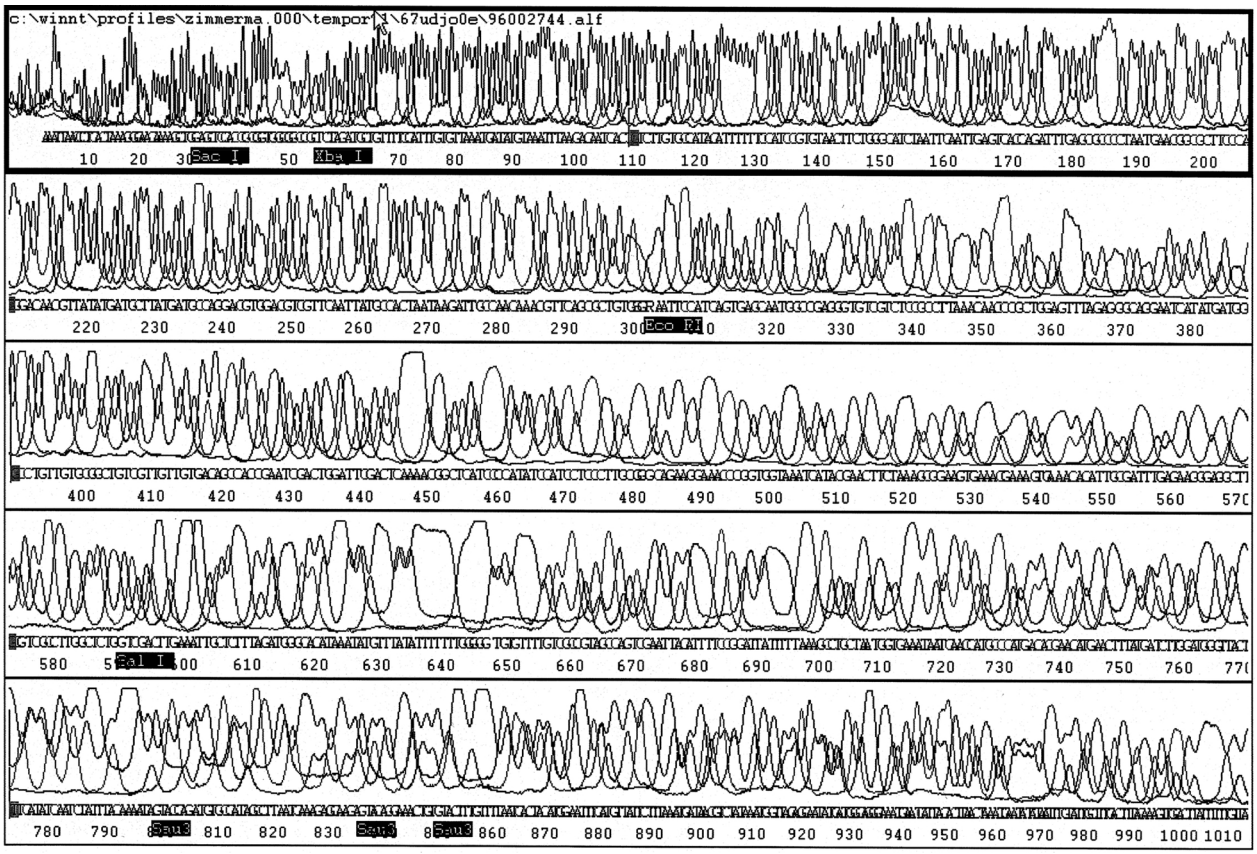


Figure 30.5 Trace (raw) data of a T7 DNA polymerase-catalyzed sequencing reaction. A plasmid was alkaline denatured, neutralized, and subjected to Sanger reaction with fluorescently labeled primers. The data were generated on an automated DNA sequencing apparatus and analyzed.

used as the original DNA sequencing enzyme, largely obsolete. In Figure 30.5, the results of a T7 DNA polymerase-catalyzed reaction are exemplarily represented with fluorescent marking margins.

The representation is different from than the top view on a sequencing gel as selected in Figure 30.4. Figure 30.5 shows a longitudinal section through the four lanes of a DNA sequence ladder. The individual bases (marks) can be reproduced in color.

Specifically, a sequencing reaction is composed with a dsDNA molecule from the following partial steps: denaturation and neutralization, primer annealing, strand synthesis, reaction stop, and final denaturation, which will be considered below in more detail.

Denaturation and Neutralization DNA sequencing reactions can be carried out both by double-stranded (ds) and single stranded (ss) DNA. They differ only in the denaturation as carried out in the first step. A denaturation in the presence of an added primer is a prerequisite for the subsequent enzymatic DNA synthesis. Single-stranded DNA is denatured by the action of heat, but only short term, and for further reaction it is brought to the optimum reaction temperature of DNA polymerase (37 °C.). In the case of the double-stranded DNA, as described here, the protocol begins with an alkaline denaturation, since heat incubation alone is not sufficient for the complete denaturation. Only combination with a strong alkaline agent such as NaOH can bring about the desired strand separation. The subsequent neutralization allows for adjustments of the reaction conditions required for the strand synthesis.

Basic Principles of Hybridization, Section 28.1

Primer Hybridization However, the ordered synthesis initiation only takes place if the DNA used is converted into a linear, single-stranded form and subsequently a primer is hybridized at the designated location. The kinetics of oligonucleotide hybridization is described by the formula derived by Lathe. It provides a connection between the hybridization time $t1/2$ in which 50% of

the oligonucleotides hybridize to a template, as well as the length, size, and the concentration of oligonucleotides:

$$t_{\frac{1}{2}} = \frac{N \ln 2}{3.5 \times 10^5 \sqrt{L \times Cn}} \tag{30.1}$$

where: $t_{1/2}$ = the hybridization time,
N = number of base pairs in a non-repetitive sequence seconds,
L = structure length of the oligonucleotides,
Cn = oligonucleotide concentration in $\mathrm{mol\,l^{-1}}$.

If one now uses the most common parameters occurring in a sequencing reaction for a primer length of 18–25 base pairs and a concentration of 10^{-7} M, hybridization times of between 3 and 5 seconds will occur.

Strand Synthesis and Pyrophosphorolysis The synthesis of DNA catalyzed by a DNA polymerase is an equilibrium reaction. The equilibrium reaction is shifted to the side of the condensation products, that is, the reverse reaction runs considerably slower than the forward reaction. With increasing reaction times and depending on the amount of DNA used, however, the reaction can run essentially backward. In these cases there may be a reduction of terminal dideoxynucleotides. This effect, which is called pyrophosphorolysis, manifests itself in disappearing sequencing bands in the sequencing gel (Figure 30.6). Pyrophosphorolysis therefore seems to be sequence-specific at selected locations in both T7-catalyzed as well as *cycle* sequence reactions. However, the removal of diphosphate from the reaction equilibrium can suppress the

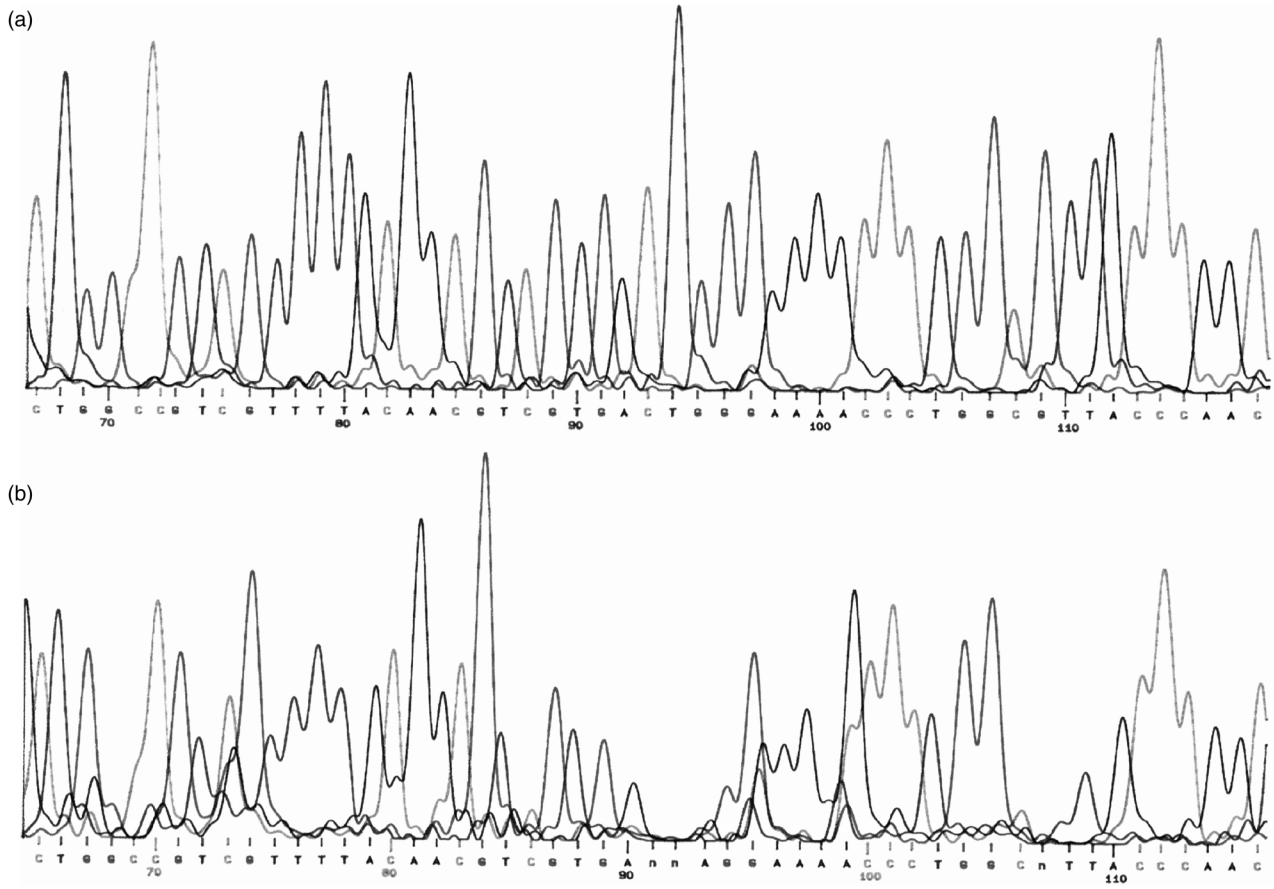


Figure 30.6 Comparison of a sequencing run with diphosphatase (a) and without diphosphatase (b). The complete lack of reaction products (position 91, 92) and the reduced signal strengths at specific positions (position 68, 86, 106) are easy to detect.

reverse reaction almost quantitatively. This can be achieved by adding a diphosphatase (pyrophosphatase), which cleaves the diphosphate into monophosphates.

Strand Synthesis and Cofactors The process of enzyme-catalyzed DNA polymerization is Mg^{2+}-dependent. It is speculated that Mg^{2+} ions are required during catalysis for the stabilization of the α-phosphate group of the incorporated nucleotide. Sequencing reactions that contain Mg^{2+} as counter-ions for nucleotides are characterized by strong variations in the signal levels of the individual sequencing bands. The partial substitution of Mg^{2+} by Mn^{2+} in T7 DNA polymerase reactions leads to homogenization of the signal intensities and thus facilitates reading of sequences, especially regions in which the resolution of the sequencing gel decreases. This effect, however, cannot be applied to thermostable DNA polymerases as Mn^{2+} inhibits the entire reaction in all the ten cases that were studied.

Strand Synthesis and Additives For additives in DNA sequencing reactions, different groups of substances are considered, such as proteins or detergents. The addition of one protein from single strand binding *Escherichia coli* (SSB) or the T4 gene 32 product could in fact stabilize single-stranded DNA structures. However, the addition shows only marginal improvements in the reaction with a high concentration necessary.

We have already discussed the combination of DNA polymerase and diphosphatase above. Further possibilities include combinations of DNA polymerases with different characteristics. In a mixture, a polymerase for the DNA brand marking can be combined with another polymerase to synthesize the sequencing products.

The reduction of background signals with the addition of DMSO in T7 polymerase reactions is attributed to its denaturing effect. The addition of formamide, and detergents such as Triton X-100, reduces the background of thermostable enzymes of catalyzed sequencing reactions.

Strand Synthesis and Nucleotide Analogs In dideoxy sequencing reactions the nucleotide analogues C7 deaza-dGTP (Figure 30.7a) and dITP (Figure 30.7b) are used. Both analogues are accepted by DNA polymerases and built into the polymerized strand. However, their derivatization prevents the formation of hydrogen-bonds. This effect is necessary in areas or structures with high GC content. Otherwise, an effect may occur that is referred to as compression. In sequence gels a zone is then observed, in which the band gap is continuously shortened and is eventually cumulated in a considerably broader zone. After this compression, the next bases occur again only after a significantly wider, empty zone. In the compression zone, more than one sequencing product can be located within a visible band. The usually appropriate spacing pattern

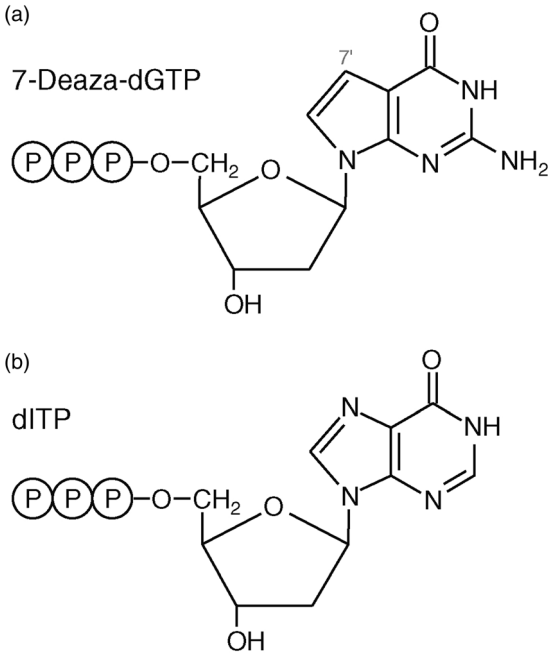


Figure 30.7 Structure of deoxynucleotide analogues in dideoxy sequencing reactions: (a) C7 deaza-dGTP and (b) dITP.

for a base is severely disrupted. This compression is caused by the interaction of highly GC-containing, complementary sections. It can result, for example, from hairpin structures, which differ drastically from regular sequences in their mobility.

Final Denaturation For exact sizing of DNA fragments a complete denaturation is necessary to prevent a sequence-dependent folding of DNA molecules or the formation of conglomerates among multiple DNA molecules and thus an uncontrollable runability. For the denaturing agent urea is usually used at a concentration of 7–8 M. Because of their polar properties, the carbonyl group and the amino groups of urea compete with the individual bases for the formation of hydrogen bridge bonds and can thus prevent the formation of base pairs. The additional use of formamide – after completion of secondary sequencing reactions and subsequent heat treatment – already results in a broad denaturation in the reaction vessel before applying the sample onto the sequencing gel. Chelation of the divalent metal ions (Mg^{2+}, Mn^{2+}) present in the reaction environment through EDTA leads to the dissolution of the DNA polymerase complexes.

Cycle Sequencing with Thermostable DNA Polymerases The information obtained with T7 DNA polymerase about the structure and function of the enzyme could be transferred to thermostable DNA polymerases in wide ranges. By way of genetic engineering it was thus possible to completely remove 3′-5′ exonuclease activity. The positive feature of T7 DNA polymerase by which it discriminates only slightly between dNTPs and ddNTPs could be attributed to the tyrosine residue 526 of the nucleotide binding point. An exchange of the corresponding phenylalanine residue at position 667 of the *Taq* DNA polymerase also allowed for a lower discrimination. Today, thermostable DNA polymerases are available that match the properties of T7 DNA polymerase in essential points, but have considerable advantages because of their thermal stability. Mixtures of polymerases, additives, and modifications that can mediate, for example, hot-start properties make cycle sequencing the method of choice.

The application of thermostable DNA polymerases allows for amplification and an additional sequencing in DNA sequencing analogous to PCR (Figure 30.8). This process is called *cycle sequencing*. Unlike PCR there is only one primer in the reaction, so it will be amplified only linearly and not exponentially. The repeated heat incubation is also sufficient for the denaturation

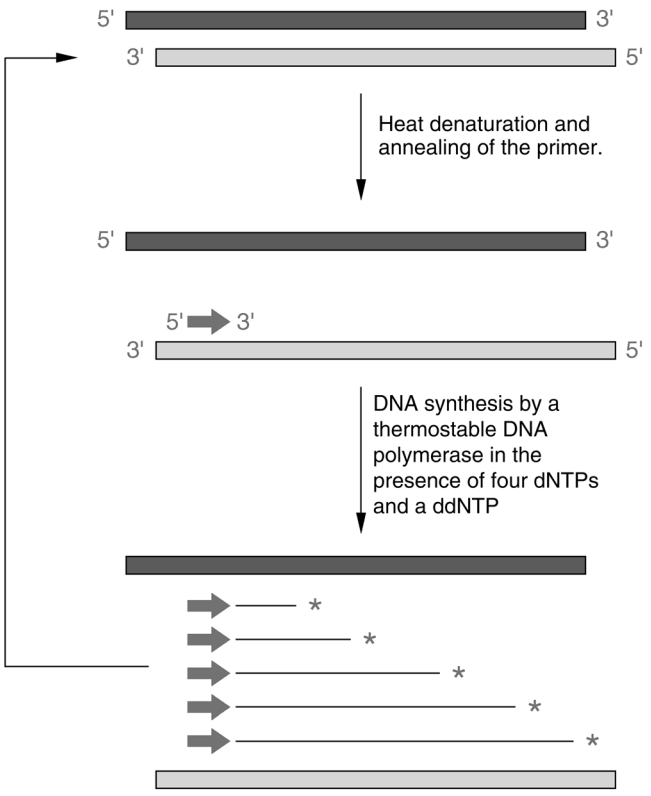


Figure 30.8 Principle of a cycle sequencing reaction. In a cyclically (about 30 times) accumulated sequencing reaction, a mixture is incubated consisting of primers, template, thermostable DNA polymerase, dideoxynucleotides, and the deoxynucleotides at 97, 60, and 72 °C. Because of the high temperatures, primer and template are thermally dissociated. At the lowest temperature occurring in the process, primer and complementary DNA section are associated, to be extended and terminated at a medium temperature through DNA polymerase. Source: adapted according to Strachan, T. and Read, A.P. (2005) *Human Molecular Genetics*, 3rd edn, Oxford University Press, Heidelberg.

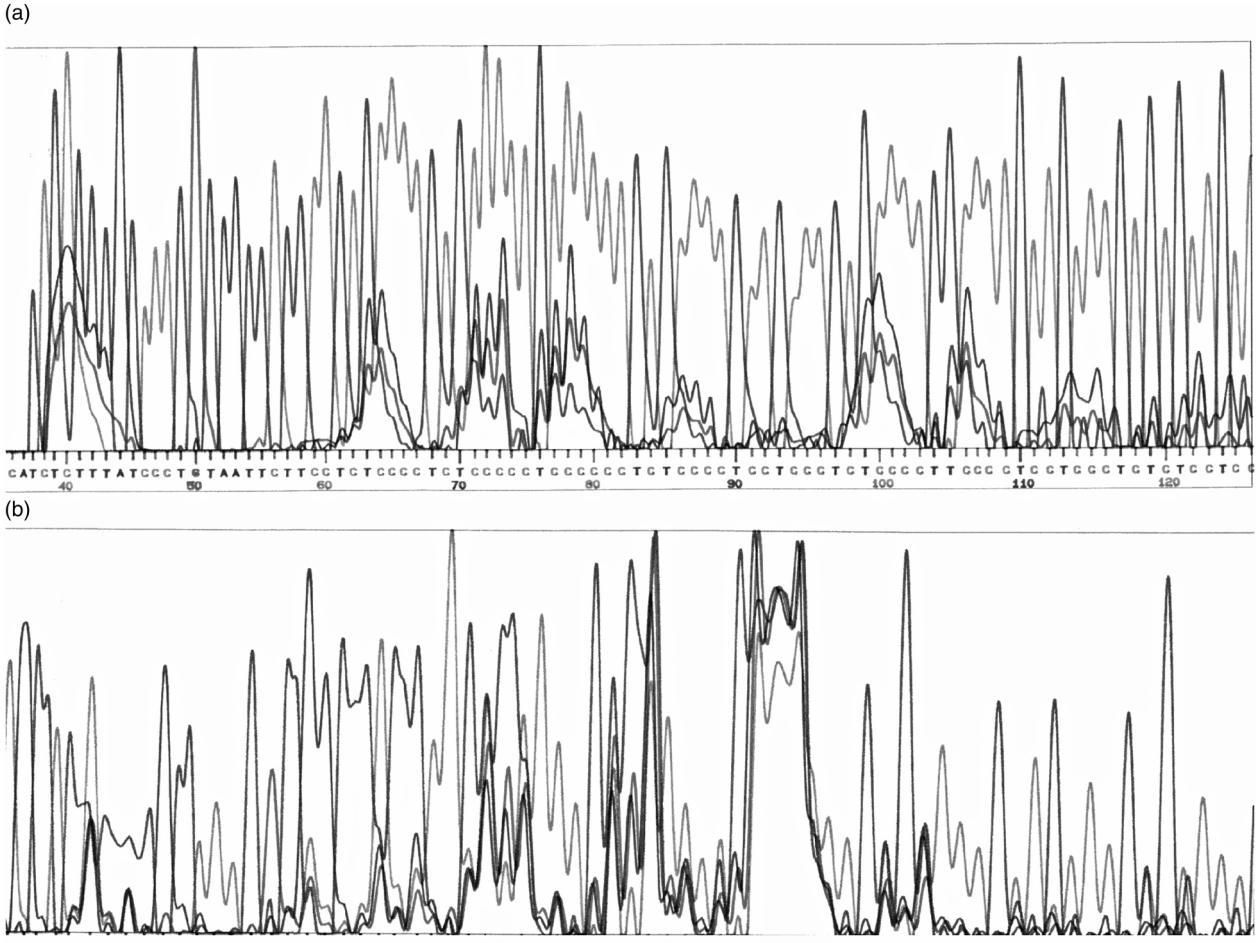


Figure 30.9 Comparison of enzyme processivity in highly repetitive sequence sections. The shown repeat has a degree of repetition of 200(!). The structure is clearly resolved with T7 DNA polymerase (a), while a thermostable DNA polymerase (b) shows an unspecific break after the first repeat.

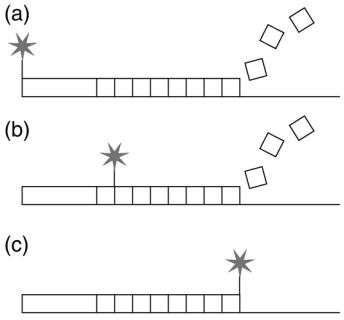


Figure 30.10 Introduction of fluorescent markings in DNA sequencing reactions: (a) labeled primers, (b) labeled deoxynucleotides and (c) marked dideoxynucleotides. In cases (a) and (b), further deoxynucleotides can be fused after installation of the selected group.

DNA Probes, Section 28.2.1

of double-stranded DNA. In a mixture of DNA templates, primers, thermostable DNA polymerase, and a dNTP/ddNTP mixture, a thermal profile, consisting of primer denaturation, primer annealing, and DNA synthesis, is repeated up to 30 times, thereby repeating a sequencing reaction up to 30 times. A correspondingly large amount of sequencing fragments is produced.

The use of genetically modified enzymes, detergents, and diphosphatase allows for the production of sequencing data that reach the quality of T7 polymerase reactions (Figure 30.9).

Despite the now improved cycle sequencing conditions there are structures, in particular repetitive sections, that cannot yet be accurately determined through these protocols. Figure 30.10 gives an example of this. Even today, DNA sequencing requires the use of different methods, since each individual method cannot cover all areas satisfactorily. Even the expensive methods of chemical fission of Maxam and Gilbert (Section 30.1.3) are rarely but still in use in difficult cases.

30.1.2 Labeling Techniques and Methods of Verification

In the following, the incorporation of marker groups in the sequencing products and their detection in automated systems is described.

Isotopic Labeling The use of radioactive isotopes in DNA sequencing is based on the fact that DNA polymerases do not discriminate between different isotopes and the incorporation into a DNA strand does not lead to changes in mobility in the gel. The isotopes ^{32}P and ^{35}S are used. The radiation from ^{32}P is of higher energy than that of ^{35}S. The exposure times in the autoradiography

that follows the gel electrophoresis can therefore be shorter with ^{32}P. However, the spatial resolution is significantly lower, due to the greater energy-related blackening areas. Therefore, especially for longer DNA sequencing runs, ^{35}S is preferred.

The label may be introduced by a radioactively labeled primer or during the sequencing reaction itself. Radioactive labeling of DNA sequencing primers is performed by phosphating with γ-^{32}P-ATP and polynucleotide kinase. The primers produced by chemical synthesis have free 5′ OH groups, which can be phosphorylated directly, without prior dephosphorylating.

The mark during the dideoxy sequencing reaction is performed by adding α-^{32}P-dATP. The isotope is incorporated into the synthesized DNA strand, which thus is detectable. Detection is carried out either by autoradiography or the use of image plates. The autoradiograms can be evaluated manually or semi or fully automatically with a digitizer or scanner. This method is now nearly outdated.

Fluorescent Labeling and Online Detection The introduction of fluorescent markings is more difficult than that of radiotracers. The fluorescent groups have a considerable size and are only marginally integrated in many cases by these enzymes due to steric conditions. If the detection groups are accepted, however, it is necessary to prevent statistical multiple incorporation. Owing to the other charge ratios and the additional mass they would inevitably lead to a change in the gel electrophoretic mobility and thus exclude a sequence determination. Labeling can be carried out with a labeled primer, internal labeling, or labeled terminators (Figure 30.10).

Fluorescein, NBD, tetramethylrhodamine, Texas Red, and cyanine dyes are used for labeling. The dyes can be linked to the appropriate components (nucleotides, amidites) and are accepted as either 2′-deoxynucleotides or 2′,3′-di-deoxynucleotides of DNA polymerases. Through laser excitation in detection systems, they are only slightly bleached and are sufficiently stable under coupling, sequencing, and electrophoresis conditions.

Energy Transfer Dyes Combinations of fluorescent dyes are also used. These so-called energy transfer group systems are based on the idea of being able to excite individual dyes for fluorescence whose excitation spectrum does not correspond to the excitation wavelength of the system or to use the emission wavelengths of a color that does not interact with other dyes used in the system (minimization of spectral overlap). The laser-induced emission of a dye is used to excite the actual fluorescent marker (Figure 30.11).

Primer Labeling The use of a 5′ fluorescently labeled primer in a DNA sequencing reaction is not critical. The unlabeled products that exist in each preparation of a primer are not visible in the analyzers and therefore do not disturb the results. A rarely occurring *self-priming* of the template DNA, caused by partial self-complementarity, would not be visible. Non-specific termination, which is generally due to inadequate reaction conditions or the structure of the DNA that is to be sequenced, can, however, be recognized as a non-readable structure. The marking of the primers generally takes place by linking a fluorescent amidite in the last step of the synthesis (at the later 5′ end). An amino link can also be added through another amidite and these are linked after the conclusion of the synthesis with the fluorescent dye. However, the efficiency of this process is less

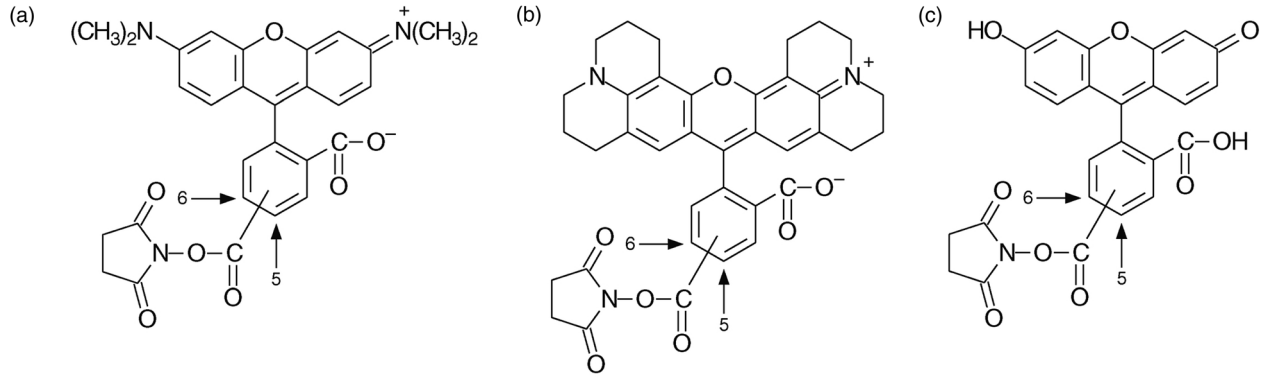


Figure 30.11 Energy transfer dyes: (a) 5- (and 6-) carboxytetramethylrhodamine succinimidyl ester (5(6)-TAMRA, SE); (b) 5-(and 6-)carboxy-X-rhodamine succinimidyl ester (5 (6) -ROX, SE); (c) 5-(and 6-)carboxyfluorescein succinimidyl ester (5(6)-Fam, SE).

than the simple linking and requires further purification steps. The labeled primer is set in the reaction similarly to a radioactive primer.

Internal Labeling with Labeled Deoxynucleotides For the internal label, fluorescently labeled 2′-deoxynucleotides are used (e.g., fluorescein-15-dATP), and a DNA polymerase. The sequencing process must run in two stages. In a first step, with a low concentration of the labeled 2′-deoxynucleotides, a maximum of one is added to each strand. A prerequisite is that the position following the primer allows the installation of this nucleotide. The concentration of marker component should be kept so low that in the next step of the reaction no further nucleotides of this type are installed. Another random incorporation would lead to uncontrolled mobility changes. The actual sequencing reaction is carried out according to the classical Sanger principle. This technique offers the advantage of using inexpensive unlabeled primers for a reaction. However, undesirable reaction products that are the result of self-priming, can also be labeled and complicate the sequencing results by overlays of multiple sequences. This method has become increasingly obsolete with the broader usage of better cycle sequencing methods.

Labeled Terminators Fluorescently labeled terminators have prevailed as a standard and offer the possibility of using non-labeled primers in DNA sequencing reactions. In the reaction mixtures, the labeled 2′,3′-dideoxynucleotides replace the completely unlabeled analogues. The marking is done through the simple incorporation of a labeled dideoxynucleotide. Another advantage is that reaction products that are not properly terminated with a dideoxynucleotide, due to the lack of fluorescence group, are not detectable. Such false, sequence-dependent reaction products are visible with labeled primers or deoxynucleotides in the system. The availability of base-specific color-labeled dideoxynucleotides allows the execution of the reaction and gel electrophoresis in one reaction vessel and a single track. The disadvantage, however, is that the DNA polymerases hardly accept these modified nucleotides, so that a high working concentration and thus a subsequent purification of the reaction products is required. Furthermore, the dyes in gel electrophoresis have different mobilities, which must be corrected by software. For self-priming products the same as was said for deoxynucleotides applies.

Duplex DNA Sequencing

Online Detection Systems In 1986 and 1987, online DNA sequencing systems were developed based on laser-induced fluorescence. Significant developments were made by L. Hood in the USA and W. Ansorge in Europe. Online detection systems consist mainly of a vertical electrophoresis system, an exciting laser, a detector, and a recording computer system (Figure 30.12). The laser is either linked transversely to the longitudinal sides perpendicular to the detector or at a certain angle from the front or rear side in the sequencing gel. The locally resolved bands of the classic method are now seen as time-resolved band patterns.

The first system of Smith, Hood, and their colleagues was based on the detection of reaction products, which were produced with the use of labeled primers in different fluorescent colors in the dideoxy method. All products of a reaction were applied to the gel on one track. The fluorescence dyes used, such as fluorescein, Texas Red, tetramethylrhodamine, and NBD, have a sufficient spectral distance to enable a secure distinction of the bases. In a further development fluorescence labeled terminators were available, which made the complex primer marker virtually superfluous. The mobility differences between the dyes used make a manual analysis of the primary data impossible, but can be corrected automatically by software programs.

The differing absorption spectra also require excitation with two different wavelengths. For observation of a gel in its entire width, a scanning mechanism has been developed. The original electrophoresis systems were based on planar gels and were later replaced by capillary electrophoresis systems, which allow a higher speed and better resolution at lower thermal load. Automated DNA sequencing systems support a much higher throughput of sequencing reactions than traditional methods (Figure 30.13). While an average of four to six reactions take place on a radioactive gel, online systems now have a payload capacity of 96 clones and read lengths up to 1000 bp per sample.

Automated Sample Preparation Automation strategies are often based on a flexible *liquid handling system*, which allows various applicative adaptations and accessories and thus a very wide range of applications with particular effectiveness. There are also devices available that have been specifically designed and optimized for a specific chemical reaction sequence. These

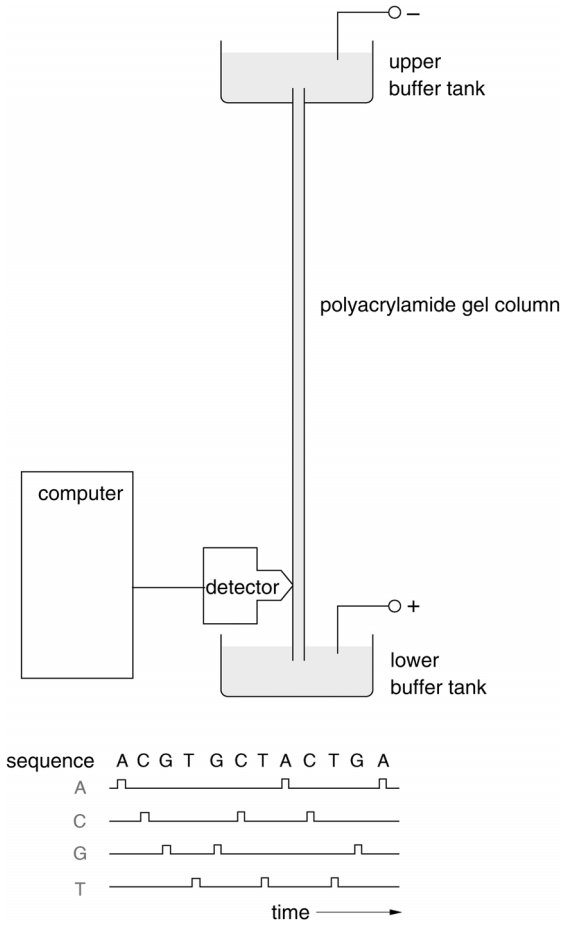


Figure 30.12 Principle of an online DNA sequencing device. A vertical gel electrophoresis apparatus is observed at the lower end of a detector. The excitating laser beam is coupled at the level of the detector into the gel (not shown here). The signals obtained from the detector are sent to an analyzing computer. The spatially resolved band patterns, derived from radioactive DNA sequencing, are replaced by a time-dissolved listed banding pattern marked by the detection finish line. Source: adapted according to Smith, L.M. *et al.* (1986) *Nature*, **321**, 674–679.

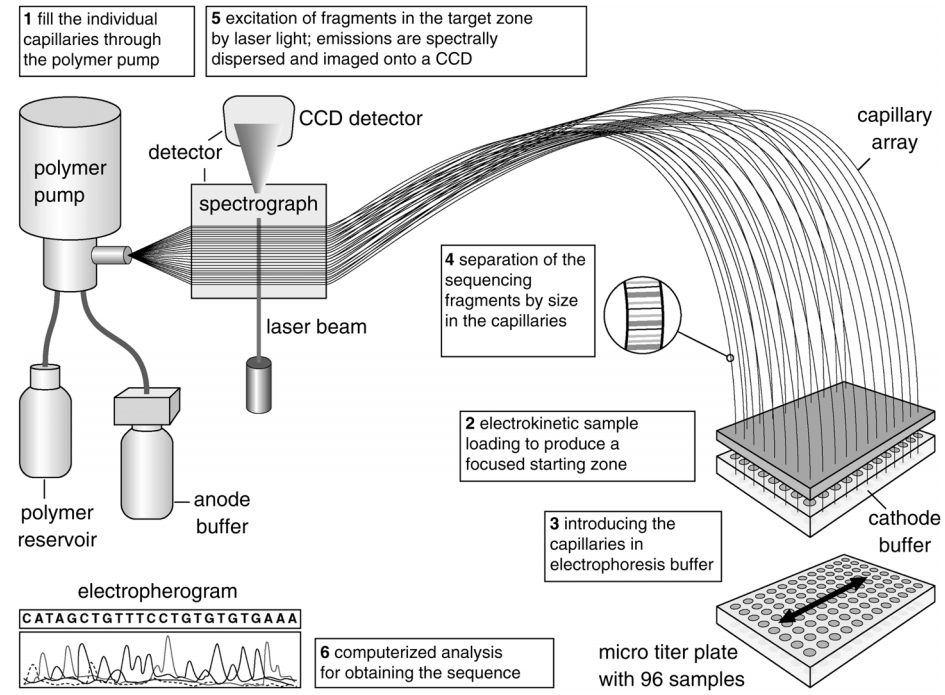


Figure 30.13 Automated capillary DNA sequencer. Source: adapted according to Perkel, J.M. (2004) *The Scientist*, **18**, 40–41 (now LabX Media Group).

different automation strategies are derived both from the different sample throughput as well as from the level of complexity of the process. Flexible automation strategies are used in the field of small and medium numbers of samples (<100 series) and used up to a medium degree of complexity. Only at high volumes or complexity degrees do specifically optimized automation concepts prove effective. Added to this the dynamics in the development of analysis and synthesis methods require a continuous adjustment and development of the strategies and thus prevent the proliferation of systems with high sample throughput, favoring instead the development of small, flexible laboratory machines.

The simultaneous processing of various samples requires a comprehensive system for documentation, quality assurance, and tracking of sample status. With the beginning of the development of LIMS (laboratory information management system) for DNA sequencing it enters into the realm of routine diagnostics and the quasi-industrial production.

30.1.3 Chemical Cleavage according to Maxam and Gilbert

Unlike Sanger, Maxam and Gilbert, instead of using enzymatic synthesis, chose the partial, base-specific chemical degradation of previously radioactively labeled DNA fragment management and also the detection of the reaction products by PAGE and film autoradiography. The method according to Sanger is now widespread due to the ready availability of specific DNA primers. However, the Maxam–Gilbert method is now only of minor importance in DNA sequencing projects. It will, though, still be utilized when the dideoxy provides non-specific termination signals due to encountered DNA structures (highly repetitive sequences, secondary structures that interfere with the polymerization, self-complementary portions).

The chemical sequencing according to Maxam and Gilbert is based on a terminal DNA labeling and subsequent partial chemical cleavage of the DNA in four independent reactions. The cleavage reactions are not specifically for only one base, but for two, depending on the reaction conditions (Figure 30.14). By combining all the tracks of a clone, a unique sequence can be determined.

End Labeling The detection of the reaction products is enabled by a 5′-terminal marking. This label can be used in both radioactive and non-radioactive form (e.g., by biotin, fluorescein). The radioactive 5′-terminal labeling of a DNA fragment is carried out in two steps. First, the 5′-phosphate groups are removed by using alkaline phosphatase. In the second step, the resulting 5′-OH group is kinased with the ^{32}P of a γ-labeled ATP. The reaction is catalyzed by a polynucleotide kinase. A 3′-terminal marking is also possible, either with a DNA polymerase

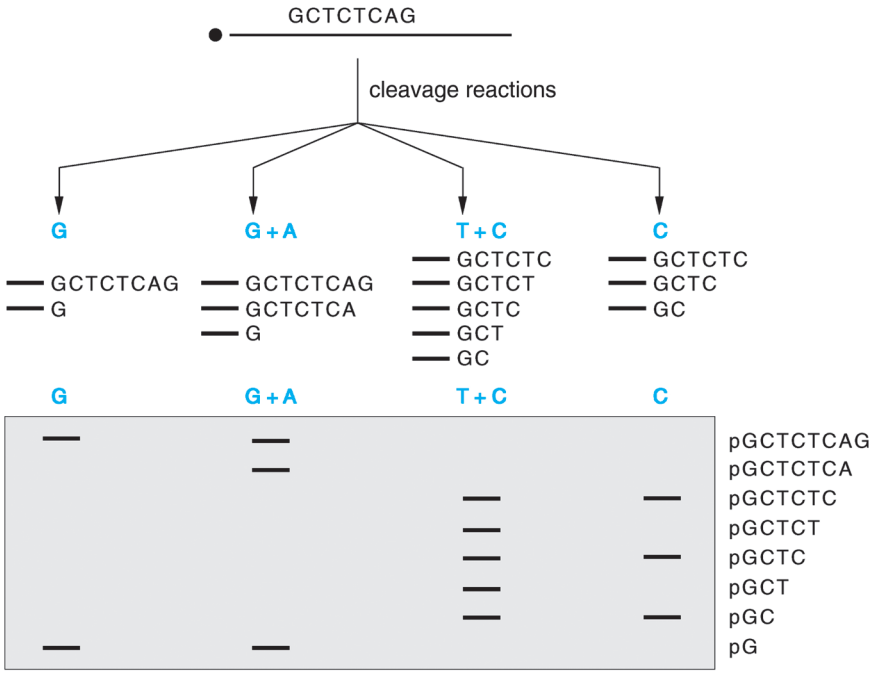


Figure 30.14 Principle of the Maxam–Gilbert method. A 5′-end-labeled DNA fragment is subjected to base-specific cleavage reactions G, G+A, C+T, and C. The reaction products are loaded in four tracks on a polyacrylamide gel. In contrast to the Sanger method, due to the non-absolute specificity of cleavage, reaction products can occur on two tracks in the same height of gel.

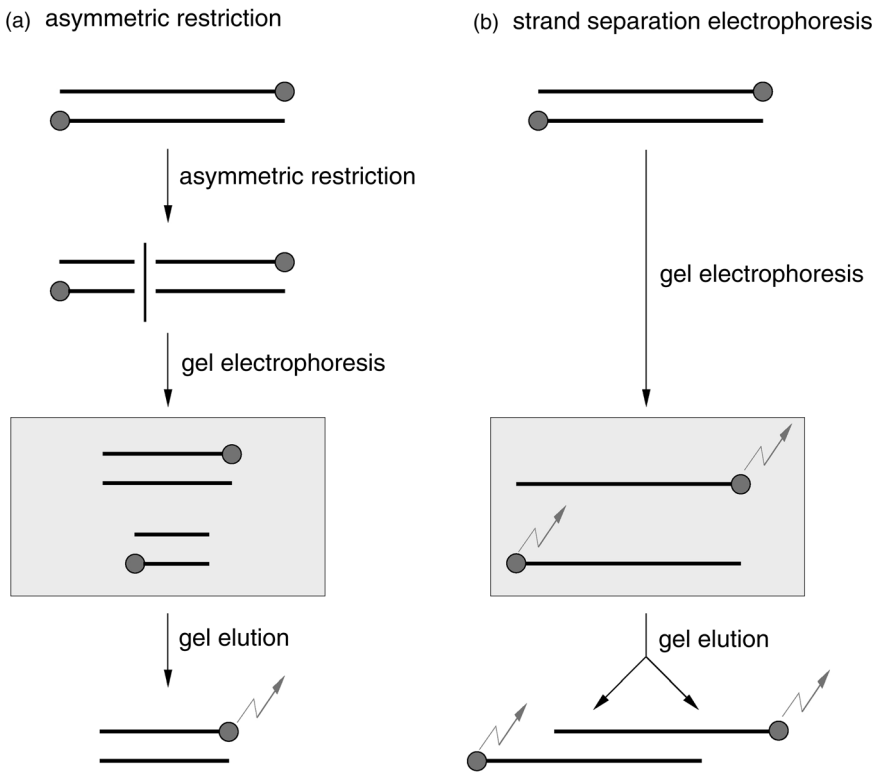


Figure 30.15 Isolation of individual, labeled DNA ends: (a) asymmetric restriction and gel electrophoresis; (b) strand separation electrophoresis.

and α-^{32}P-dNTP, α-^{33}P-dNTP or – in special cases – by terminal transferase and/or α-^{32}P-ddATP/α-^{32}P-3′-dATP (cordycepin triphosphate).

A double-stranded DNA fragment has two 5′-phosphate groups, so that the marking reaction marks both ends. In the following specific cleavage reaction both DNA strands would be sequenced simultaneously. The resulting sequences would overlap and thus would not be readable. For this reason, only a labeled strand of DNA may be present in the subsequent cleavage reaction. For the isolation, two different strategies can be applied (Figure 30.15): (a) The selected product is cleaved in an asymmetric manner with a restriction endonuclease, so that a larger and a smaller sub-fragment are released. The latter can be easily separated in a gel electrophoresis, subsequently isolated from the gel and subjected to the cleavage reactions. (b) The two ends can also be separated by a strand separation electrophoresis. Particularly convenient is the different mobility of complementary DNA strands during electrophoresis. In contrast to the first method, both strands can be completely recovered and sequenced, in separate reactions.

However, a unique fluorescent or radioactive label can also be inserted considerably more easily by using a labeled primer: by using a fluorescence-labeled primer and an unlabeled primer in a PCR reaction, a one-side labeled product can be produced, which can be sequenced directly after the buffer exchange.

Cleavage Reactions The cleavage reaction runs in two or three stages. In a first step, the bases are chemically modified. These modifications often lead to base splitting (e.g., when treated with formic acid, see below). Then, the DNA backbone is broken at the location of the modified base by treatment with alkali, whilst *in toto* a nucleoside (the modified base and the deoxyribose) is eliminated. In any case, a 3′- and a 5′-phosphate remain. The following modification/cleavage reactions have been described:

- Limited treatment of DNA with dimethyl sulfate. For dsDNA, this results in the methylation of the N3 atom of adenine and at a four- to ten-times higher chance of methylation of N7 guanine bases. Depending on the further treatment, different specificities of subsequent cleavage can be achieved:
- **G:** The methylated DNA is treated directly with hot piperidine. This leads to the specific cleavage of methylated guanine residues (Figure 30.16a).

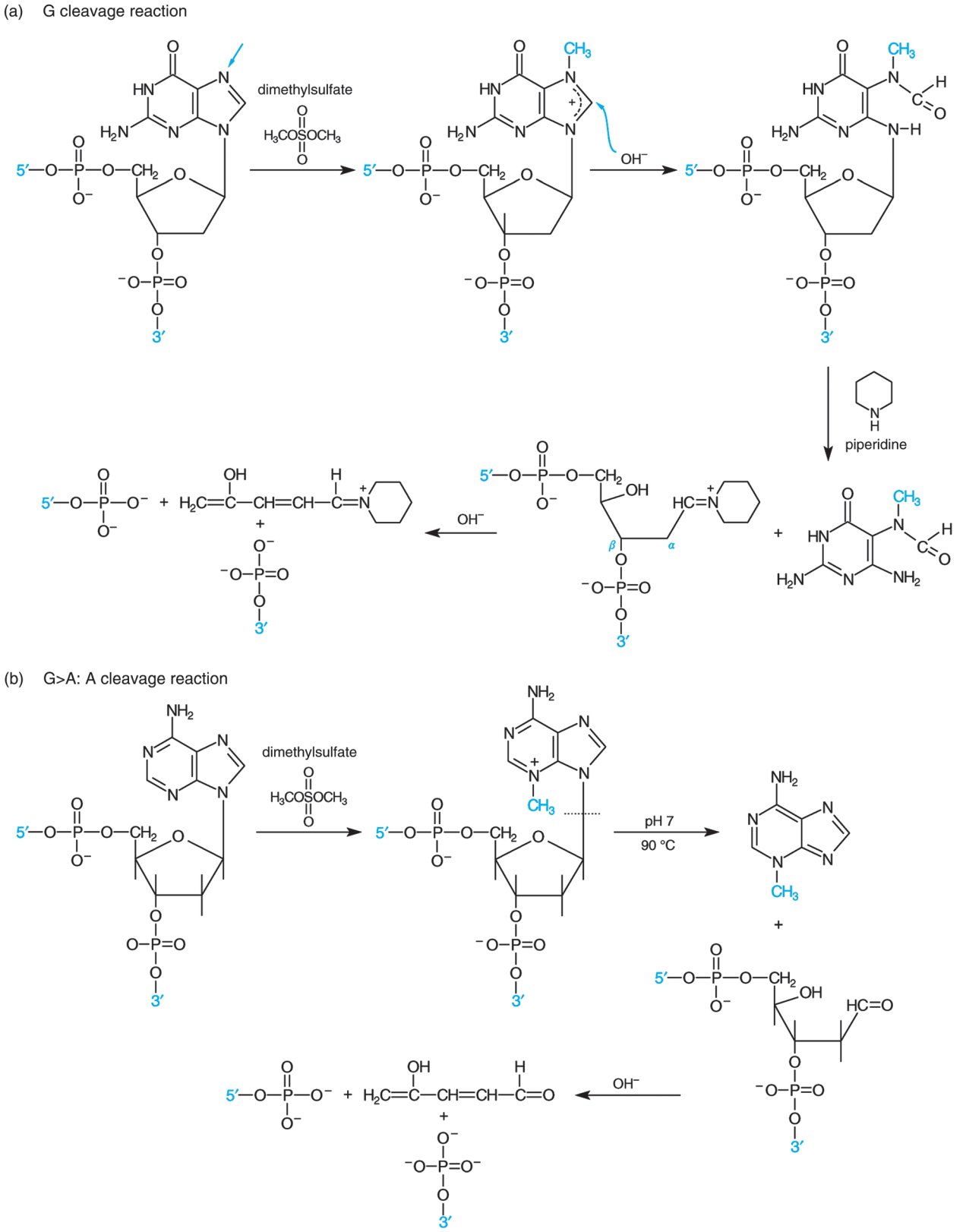


Figure 30.16 (a) *A G-cleavage reaction:* After the methylation of N7 of the guanine residue by dimethyl sulfate, a basic attack on C8 results in a ring-opening. The added piperidine causes the release of the base and at the same time β-elimination of both phosphates. (b) *G> A: A-cleavage reaction:* After methylation of the adenine residue N3 with dimethyl sulfate, the N-glycosidic bond is cleaved. The added piperidine leads to the elimination of the base and simultaneous β-elimination of both phosphates.

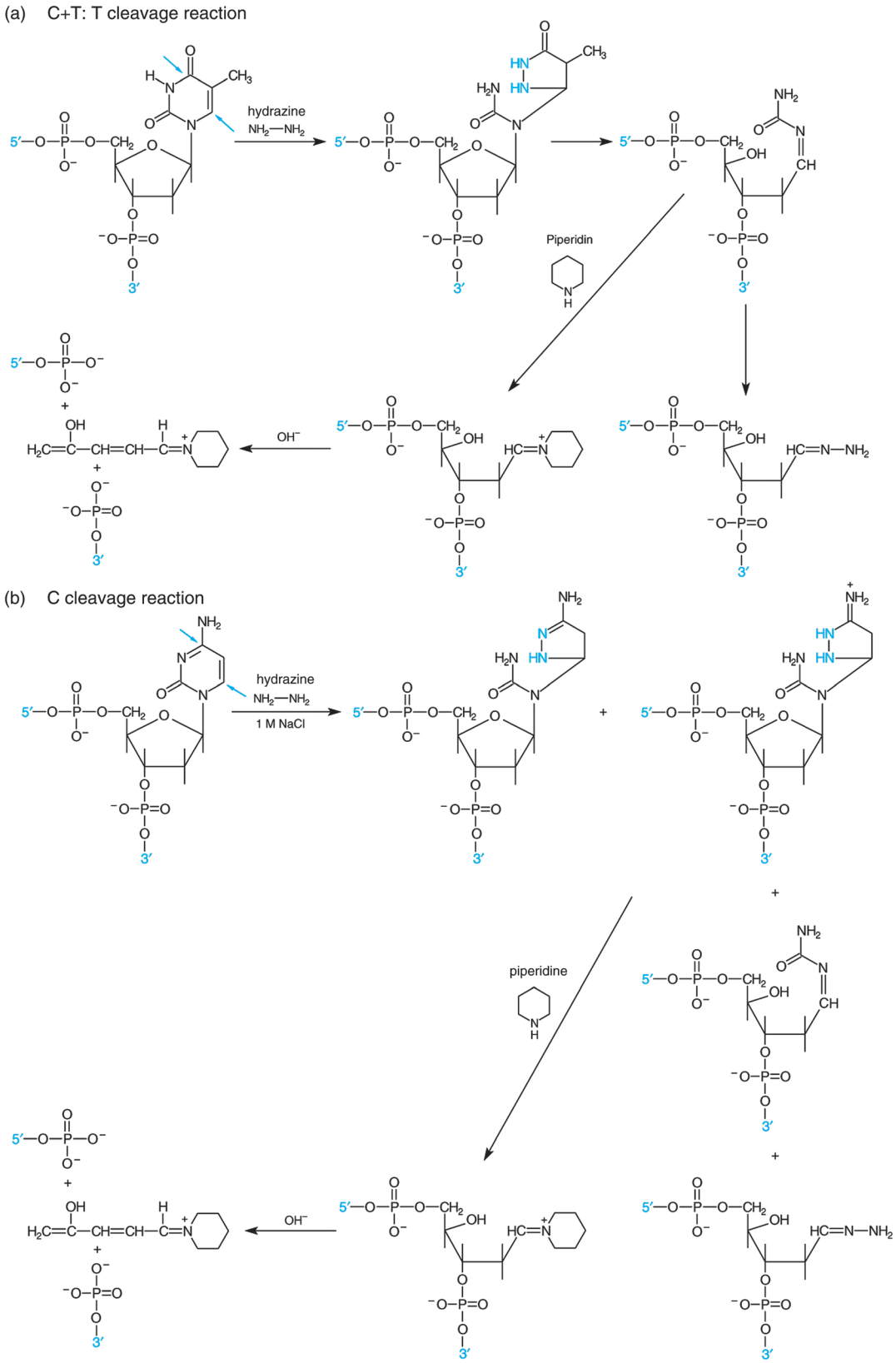
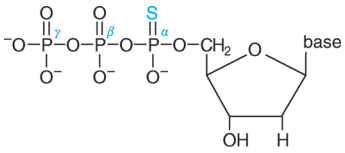


Figure 30.17 (a) Thymine- and cytosine-specific fission reaction. Hydrazine leads to a ring opening of the base between C4 and C6 and their fission. The added piperidine leads to the simultaneous β-elimination of both phosphates. (b) C-cleavage reaction. Hydrazine opens the ring between C4 and C6 and splits off the base residue. The added NaCl suppresses the reaction of thymine. The added piperidine leads to elimination of the base and simultaneous β-elimination of both phosphates.

- **G > A:** The methylated DNA is first heated at neutral pH for 15 min at 90 °C. This leads to the elimination of all methyl-A and all methyl-G. The A-depurination occurs four- to six-times faster than the G-depurination. Because of the low adenine methylation degree, however, the G-depurination outweighs the A-depurination. Hot alkaline treatment (piperidine or NaOH) of the partially depurinated DNA therefore leads to a strengthening of the G-cleavage and weakening of the A-cleavage (G > A-pattern). The A-part of the G > A-reaction is shown in Figure 30.16b. Single-stranded DNA is also methylated at the N1-atom of adenine and at C3 of cytosines. Methylation on N1 of adenine does not lead to cleavage. At C3 methylated cytosines are cleaved and are visible as bands with about 1/10 to 1/5 of the intensity of the G-bands.
- **A > G:** The methylated DNA is treated at 0 °C with dilute acid (e.g., 0.1 M HCl) for 2 h. This leads to a preferential hydrolysis of the glycosidic bond of methyladenines. Hot alkali treatment of the depurinated DNA thus results in a stronger A- and in a weaker G-fission (A > G pattern).
- **A > C:** To achieve this specificity, the DNA is treated with strong alkali (1.2–1.5 M NaOH) at 90 °C for 15–30 min. This opens the adenine and cytosine rings. Hot piperidine treatment leads to the elimination of these bases and to A > C cleavage patterns.
- **A + G:** Limited treatment of DNA with formic acid leads to unspecific depurination. Alkali treatment generates the A = G or A + G pattern.
- **C + T:** Cleavage at the pyrimidine bases is performed by modification with aqueous hydrazine and subsequent cleavage with alkali. The chemical reactions of the thymine part of this combined reaction can be seen in Figure 30.17a. Piperidine reacts with all glycoside products generated by hydrazine.
- **C:** Inclusion of 1–2 M NaCl in the previous reaction suppresses the reaction of hydrazine with thymine. The steps of the cytosine reaction can be seen in Figure 30.17b. Piperidine reacts here with all the glycoside products of the hydrazine.

Except in the case of the G-reaction, NaOH (0.1 M) or piperidine (1 M) can be used as alkali. Piperidine is preferred because it can then be removed very easily by evaporation. When using NaOH, the DNA must be precipitated with ethanol.

More base-specific reactions of DNA are presented in Chapters 31 and 32. However, for the methods discussed here, they are not equally important for the determination of the DNA sequence. Basically, the reactions G, G+A, C+T, and C are sufficient to be able to read a sequence clearly.



2′-deoxy-5′-*O*-(α-thio)-triphosphate

Figure 30.18 Structure of α-thionucleotides.

Solid Phase Process The chemical cleavage could be greatly facilitated by the development of a solid phase process and its adaptation to the automatic DNA sequencing that works with fluorescence. After binding of the DNA template to an activated membrane surface, all cleavage reactions and required washing procedures up to the elution of the reaction products can be performed on a solid surface.

α-Thionucleotide Analogues The incorporation of α-thionucleotides (Figure 30.18), for example during a PCR reaction, allows a simple chemical cleavage reaction with 2,3-epoxy-1-propanol and has the additional advantage that a DNA strand is not attacked by exonuclease activities; compared to the standard Maxam–Gilbert method, 5′-O-1-thiotriphosphate shows good incorporation characteristics with major DNA polymerases. However, the band patterns generated only allow a reliable sequence determination in short areas.

Multiplex DNA Sequencing Multiplex DNA sequencing (Figure 30.19) is based on the above chemical cleavage of DNA. While in the above case detection can take place immediately, it requires sophisticated hybridization patterns in multiplexing. The goal is the simultaneous performance of 50 sequences in a reaction, thereby minimizing the workload of the elaborate cleavage reactions and gel operations. The 50 DNA fragments to be sequenced are cloned in 50 different vectors. These vectors differ in the fragments of left and right flanking linker sequences. This vector-specific oligonucleotide follows a standardized restriction site. After the excision of the DNA fragments they are enclosed by the above-mentioned sequence regions. All fragments are united in a cleavage reaction, separated by gel electrophoresis, and transferred to a nylon membrane (blotting). The reaction products are detected by successive hybridizations of the

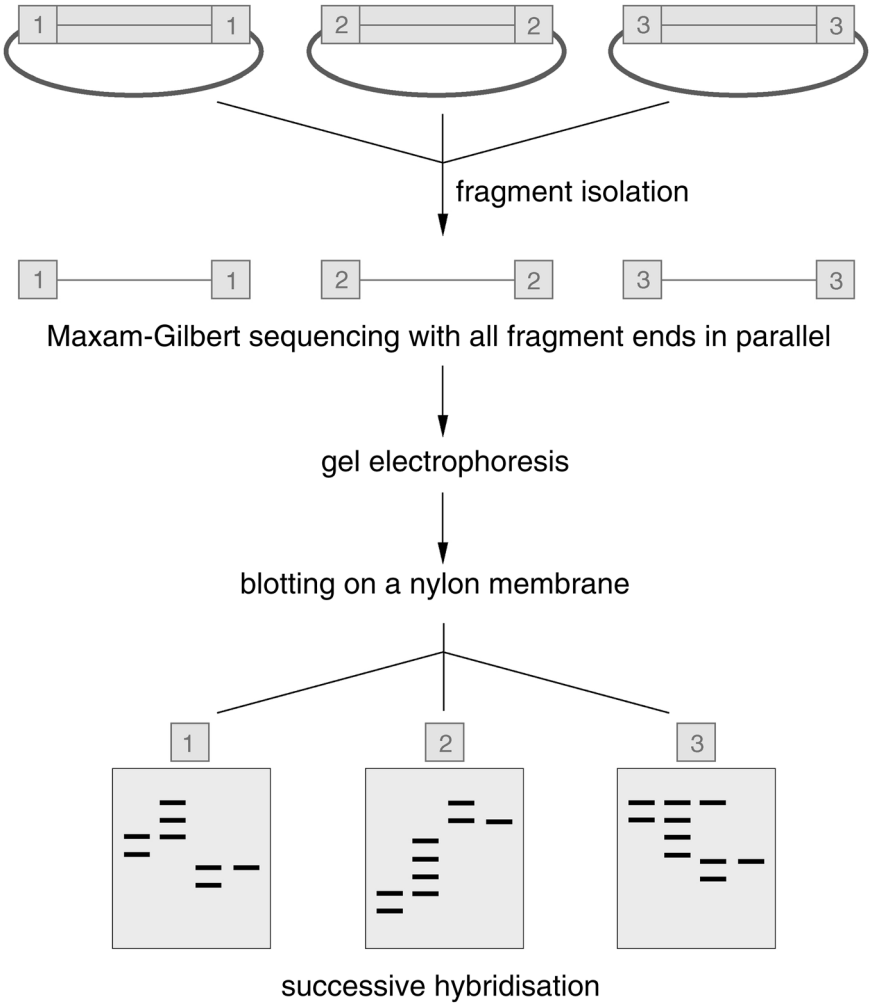


Figure 30.19 Multiplex DNA sequencing.

fragments with labeled oligonucleotides that are complementary to the flanking regions. In this way, a filter can be recycled up to 50 times. The variant of the *multiplex walking* dispenses with the elaborate cloning at the beginning of the process, and instead fragments a cloned DNA and then essentially follows the above procedure. The first hybridization begins at the start point specified by the vector. From the obtained sequence a new oligonucleotide is then synthesized and hybridized again. By repeating the process, new starting points are generated that make it possible to travel along the fragment (primer walking).

RNA Sequencing Typically, mRNA is not sequenced directly, but converted by reverse transcription into cDNA and enzymatically sequenced by Sanger's method. For sequencing of rRNA, chemical cleavage is sometimes used. In chemical RNA sequencing, we are talking about a chemical cleavage by a Maxam and Gilbert method analogue. Owing to the lower stability of the phosphodiester bond and other chemical compositions of RNA, modified cleavage reactions are used. A 3′-terminal marked RNA fragment is subjected to four parallel batches of base-specific modification reactions and the RNA strand is split with aniline instead of NaOH or piperidine:

- **G-reaction:** base methylation with dimethyl sulfate, followed by reduction with sodium borohydride and cleavage of the modified RNA strand in the ribose bond with aniline.
- **A> G-reaction:** ring opening of the bases with diethyl dicarbonate (diethyl pyrocarbonate) at N7 and subsequent strand cleavage with aniline.
- **U-reaction:** Treatment with hydrazine results in cleavage of the base by nucleophilic addition to the 5,6-double bond. The modified RNA strand is cleaved by aniline.

- **C> U-reaction:** Treatment with anhydrous hydrazine in the presence of NaCl leads to preferred release of cytosine. Again, the strand is cleaved by aniline.

30.2 Gel-Free DNA Sequencing Methods – The Next Generation

Gel-free sequencing methods brought in a new wave of sequencing technologies starting in 2005 with a version of pyrosequencing technology named 454. One particular feature of this and the following technologies is that they produce large amounts of sequencing data, but it is made up of generally much shorter reads than the standard Sanger sequencing method. To obtain the mass of sequencing data that characterizes the Next generation sequencing (NGS) wave, all the techniques rely on a high level of parallelization of their sequencing method. Massively parallel sequencing (MPS) is another term used for the NGS sequencing.

With the start of the NGS wave it became apparent that sequencing of individual genomes was within reach, with all the impact that this may have on health care and personalized medicine. As a result many different NGS methods have been proposed, and following on from them several attempts have been made to commercialize various methods (for a review of methods proposed and the relevant time line see Reuter, J.A., Spacek, D.V., and Snyder, M.P. (2015) *Mol. Cell*, **58** (4), 586–597). As a result of the intense method development and rapid commercialization many of the newer sequencing technologies are referred to by their commercial name – that is, Illumina sequencing and not the underlying method – sequencing by synthesis.

The gel-free methods eliminate gel electrophoresis as the throughput and resolution limiting factor in DNA sequencing, but cannot hide their similarities to the Sanger sequencing method. Many are based on the incorporation of labeled nucleotides that have the label removed after detection and are converted into an actively extendable polynucleotide string. The detection occurs *in situ* bound to a solid surface instead of a measured end point in a gel. Therefore, it is possible to achieve an extreme size reduction in the reaction volume and the related detectors. The use of activated surfaces makes a dense packing of the reacting molecules possible and thereby a huge parallelization of the sequencing and a huge sequence output. As almost every single incorporation event is detected (HiSeq, 454), the throughput is mainly governed by the microfluidics of the system in question – how quickly can the sequencing cycles be captured versus the buildup of fluorescent noise in the system with time. However, there are already new systems available that are mainly governed by the progressivity of the polymerase (PacBio RS). The next generation of sequencing devices, those that use nanopores or other methods, are capable of measuring single molecules and are already to some extent available.

Since the next generation of sequencing devices came to the market the amount of sequencing data has expanded hugely, as has the size of sequencing projects that are being attempted (e.g., 1000 genomes and 1000 cancer genomes). The rate of sequence data production continues to increase with systems such as the x10 from Illumina promising $1000 dollar human genomes. Further increases in the sequencing output are predictable, however, and the data produced from the latest sequencing instruments can measure over a terabase produced in less than 4 days. This has meant that not only are there new opportunities appearing for biological research but also new challenges appear dealing with the large amounts of data produced. New methods in bioinformatics for analysis of the data such as aligning large numbers of small sequences have had to be developed and IT infrastructure that holds and processes these data has had to be expanded. NGS data analysis can last for days or weeks even on a powerful cluster. It can have very high memory requirements (1 terabyte of RAM) and the transfer of 100 GBs of result data can push the boundaries of conventional WAN (wide area network) technology.

In some aspects the preparation of probes has become easier, the classic cloning steps are removed that can lead to an unwelcome bias, although other biases are added such as PCR sensitivity to base content. In other aspects the process does lead, however, to a much greater effort in library preparation. The available systems have a detection level of around 25 000 molecules and therefore require amplification of the sequencing product to be able to detect it. The sequencing methods are also very dependent on the size distribution of the molecules to be

sequenced, therefore shearing and accurate size selection of the input material is required to reduce the size of the starting molecules to a few hundred bases long.

30.2.1 Sequencing by Synthesis

Classic Pyrosequencing Through analysis of the byproducts of a polymerization reaction one can determine the nucleotide sequence (pyrosequencing, Figure 30.20). As mentioned above, every polymerization reaction results in a free diphosphate and these can be detected by another chemical reaction. The enzyme sulfurylase catalyzes the conversion of the diphosphate into ATP, which when hydrolyzed with luciferase results in luciferin and oxyluciferin.

The polymerization can be detected by adding single nucleotides to a polymerization mixture in a reaction chamber. A successful polymerization produces the expected emission of light. Through successive cycles of nucleotide addition, detection of polymerization by emitted light, and removal of the reaction products, the sequence of a nucleic acid can be determined. Light is only emitted when the correct nucleotide for the sequence is added. The presence of homopolymer stretches results in a larger signal based on the number of nucleotides incorporated although the increase in signal cannot be relied upon to be a linear representation of homopolymer length. In general, stretches of more than eight identical nucleotides cannot be reliably quantified. Each cycle time takes a few minutes per base and read lengths of around 60 bases are achievable, although in this form pyrosequencing has been generally regarded as a mini-sequencing technique.

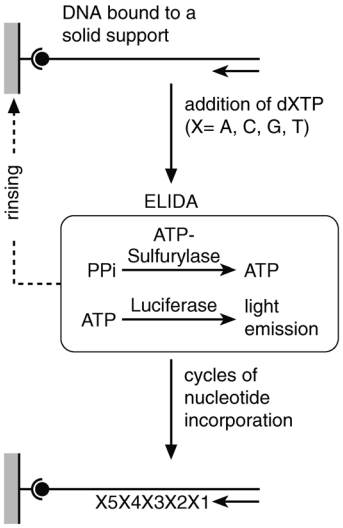


Figure 30.20 *Principle of pyrosequencing:* An immobilized DNA probe is incubated with a reaction mixture containing only one nucleotide (A, C, G, or T) in cycles. If the correct nucleotide is currently present in the mixture it is incorporated in the polymerizing nucleotide chain and a diphosphate is released which is converted into ATP by ATP-sulfurylase. This is in turn used by luciferase to convert luciferin into oxyluciferin along with the emission of light. This light emission is used to measure the incorporation of the base or bases in this cycle. Source: Ronaghi, M. *et al*. (1996) *Anal. Biochem.*, **242**, 84–89. With permission, Copyright © 1996 Academic Press. All rights reserved.

454-Technology (Roche) The 454-system was the first second-generation MPS device to the market and uses pyrosequencing as its basis although the sequence read lengths are longer than in classic pyrosequencing and the high level of parallelization of the process gave a huge increase in sequence data output at the time. The sequencing process (Figure 30.21) is made up of three steps:

1. Creation of a library of DNA molecules in the range 300–800 bp. The starting material that is longer than the acceptable range, such as genomic DNA, is sheared by nebulization, purified, the overhanging ends repaired and then phosphorylated (Figure 30.21a, 1). Afterwards two adapters, A and B, are ligated to the target molecules (2). The adapters are each 44 bp long themselves and are not phosphorylated but carry the target sequences for amplification and sequencing. A 4 bp long key sequence identifies the library sequences to the system and facilitates the calibration of base recognition and calling. This step is left out when sequencing PCR products as the necessary adaptors should have been added during the PCR amplification itself. The B adaptor has an extra 5′-biotin modification. This facilitates a cleaning step using streptavidin-covered paramagnetic beads (3). The library fragments without biotin are simply washed away. The gaps between the adapter and the fragment are filled by a strand-displacement DNA polymerase catalyzed reaction. During the cleaning procedure alkali denaturing of the one sided modified fragments and occasionally the unmodified and complimentary strand are set free for further steps, while double labeled fragments spontaneously reattach under the chosen conditions.
2. The DNA being sequenced has to undergo clonal amplification to give a clear signal above the detection system's lower limit (4, Figure 30.21a). A bead dilution is performed so that ideally only one capture bead binds to one DNA molecule. The beads carry matching capture primer, allowing them to bind to a DNA fragment. The beads are then mixed in an oil and water emulsion, where an emulsion-PCR reaction (emPCR) takes place and the sequences are amplified (5). A further biotinylated primer is used to allow downstream purification of clonally amplified sequences. After the PCR reaction the emulsion is broken and the beads binding DNA, byproducts, and empty beads are freed. Only the beads with amplified sequences are bound by streptavidin beads and magnetically filtered out of the mixture. Through the breaking of the emulsion and following alkali denaturing a bead is produced that has a bound, single-stranded DNA molecule that can be sequenced.
3. For DNA sequencing (Figure 30.21b) the sequencing primer is hybridized to the bound template in a reaction mixture containing sequencing primers, DNA polymerase, and required cofactors. The mixture is transferred to a pico-titer plate with the aim of having just one bead occupying one well of the plate. Each well has a diameter of 44 µm and has a single

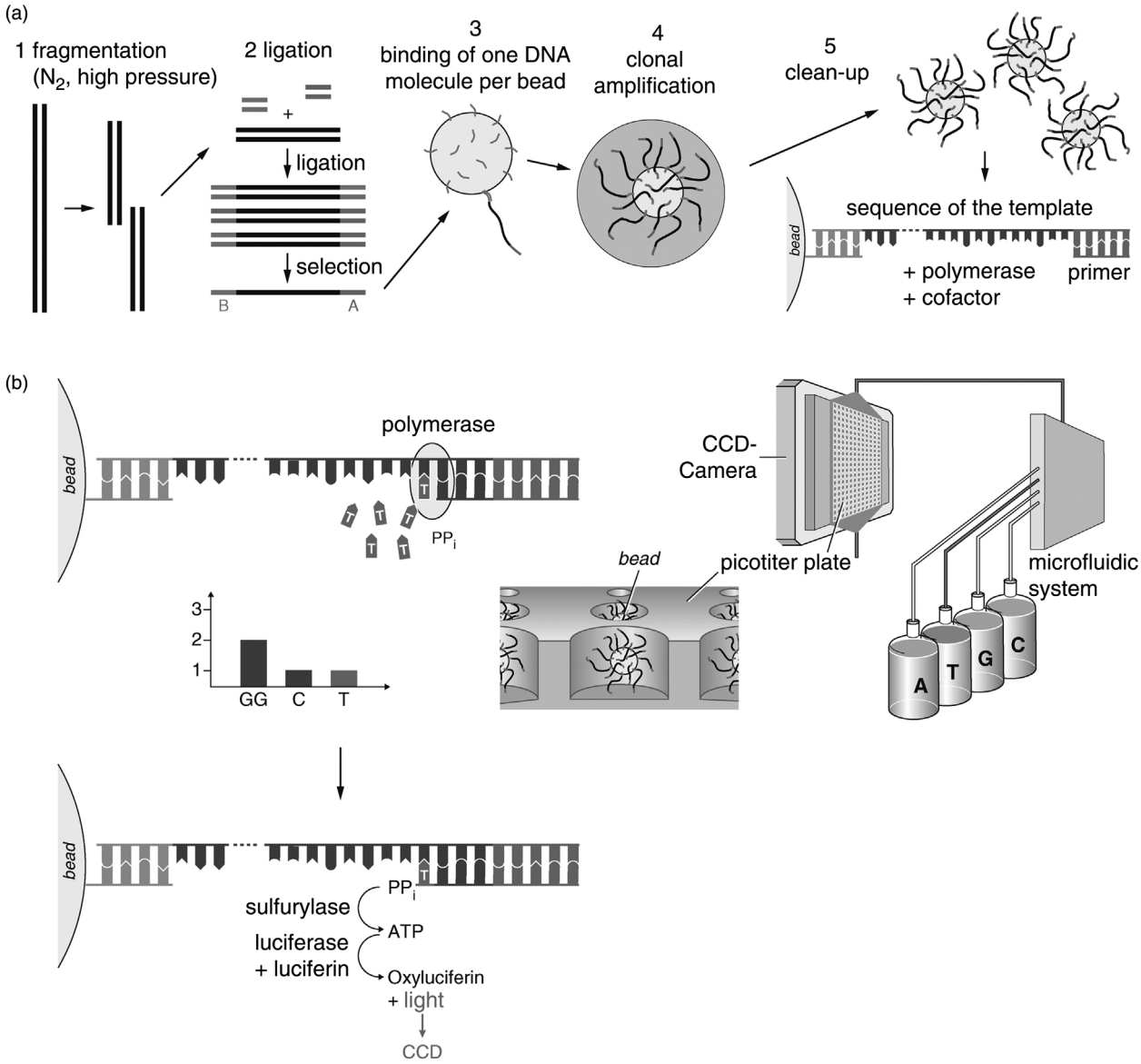


Figure 30.21 454-System Workflow: (a) Library preparation: 1, fragmenting the nucleic acid to a length between 300 and 800 bp; 2, ligation of A and B adaptors; 3, binding of template to a capture bead; 4, bead after clonal amplification; 5, bead after the emulsion is broken and clean-up. (b) Bead bound pyrosequencing. Source: adapted after Roche Diagnostics Corporation.

Phred: The calculation of Phred quality scores dates back to the automatic gel sequencing methods of the 1990s. The value is based on the error probability for each base in a read and is calculated using the following formula:

$$Q = -10\log_{10}P$$

A Phred score of 40 represents an error probability of 1 in 10 000. The Phred scores are presented as a string of characters where the ASCII encoding value for the character minus a particular offset, 33 in most cases, is the calculated Phred score.

engraved glass fiber within the plate housing assigned to it. The fiber in each well transfers its signal to a point on a CCD sensor, allowing the pattern of sequencing to be detected. The dimensions of the well prevent more than one capture bead occupying it. Added enzyme beads contain the two enzymes luciferase and sulfurylase, which generate emitted light in the presence of PPi. The process follows the above-mentioned classic pyrosequencing; however, 400 000 reaction wells are simultaneously active and the achieved read length is significantly longer.

During the runtime the detector pictures are analyzed and the signal intensities calculated from the pixel data. These data are reduced to give the intensity values assigned to the particular positions on the pico-titer plate. The series of single pictures is used to finally calculate the quality values for each base (Phred encoded probability of error) and the read sequence is presented as a fluorogram. Yields of the 454 high-throughput variant of this system are around 700 MB with an average read length of 700 bases in around 23 h. The 454 system was a major breakthrough in the world of sequencing technology.

Illumina Sequencing by Synthesis (HiSeq, MiSeq, NextSeq500, ×Ten) The Illumina technology is based on the integration of fluorescent markers with a reversible terminator dNTP and the result of sequencing result is based on the optical detection of incorporated fluorescent markers through sequencing cycles.

As with other sequencing methods a library preparation step is required here as well. Ultrasound is used to produce a set of DNA fragments of the correct length, targeting 200–500 bp. A mixture of T4-DNA-polymerase and Klenow fragment removes the 3′-overhangs and fill up the 5′-overhangs. T4-polynucleotide kinase phosphorylates the blunt ends of the fragments and incubation with a 5′-3′-exo-deletion mutant and dATP leads to a simple adenylation of the ends. Finally, both the differing end adaptors possessing the necessary 5′T overhang can be ligated (1 in Figure 30.21). The P5 adapter possesses a region complementary to the sequencing primer and the P7 adaptor attaches a complementary sequence to allow binding of the fragment to the flow-cell where sequencing will actually take place. Size selection and separation from unligated adaptor sequences are performed by gel electrophoresis followed by gel elution, or magnetic bead purification, and PCR can be used to further enrich the sample being sequenced.

For the DNA to be actually sequenced it must be bound to the sequencing flow-cell and clonally amplified to allow detection. This is done through a process known as clustering (Figure 30.22).

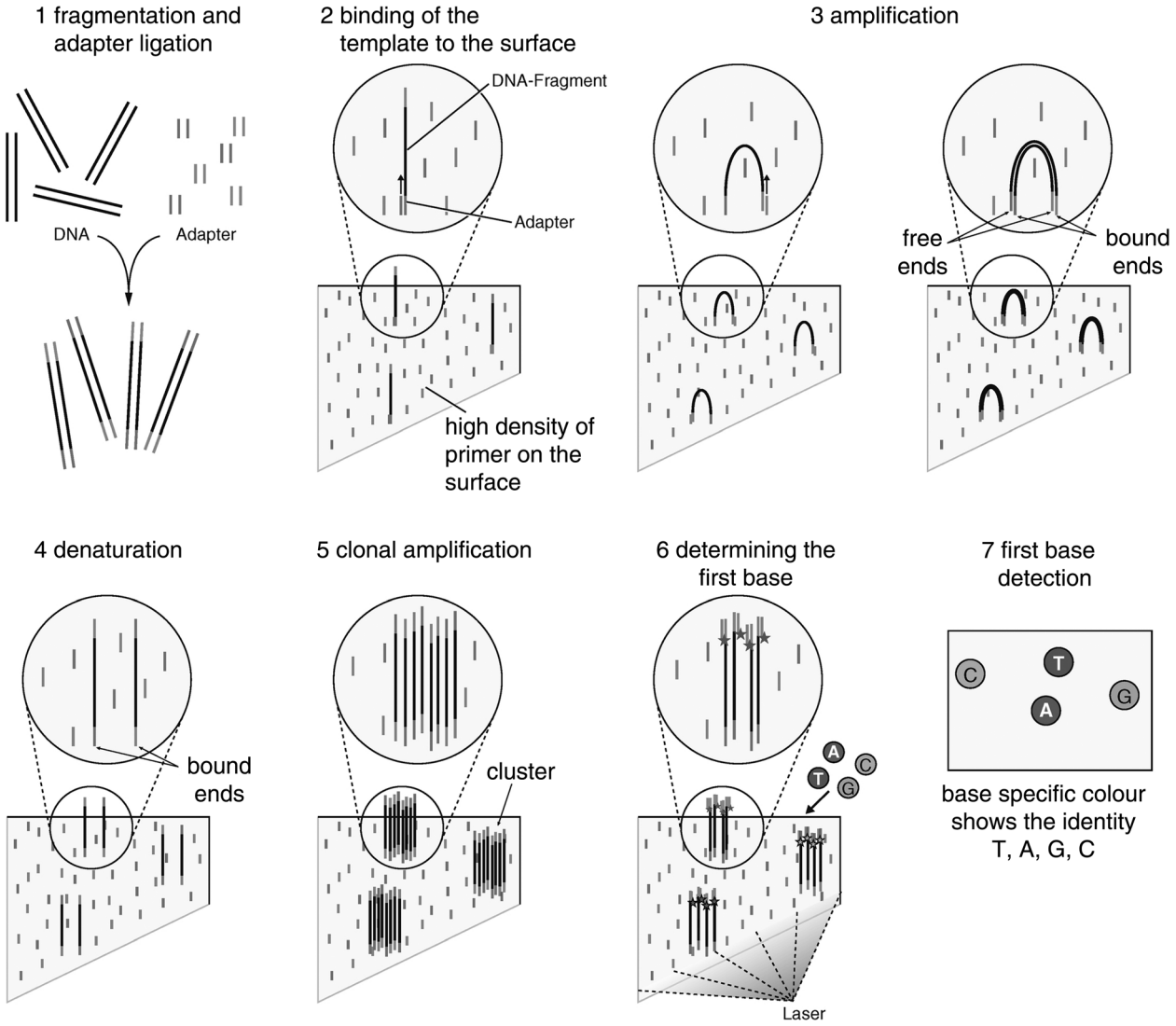


Figure 30.22 Illumina sequencer workflow: 1, fragmentation of the nucleic acid and adapter ligation; 2, loading of the template library on to the flow-cell; 3, strand initialization; 4, template preparation by denaturation; 5, clonal amplification; 6, incorporation of fluorescently labeled reversible terminators; 7, detection of the incorporated bases using the scanned images.

Clustering occurs on the transparent flow-cell where the sequencing takes place. The flow-cell can have up to eight channels (HiSeq) allowing eight separate DNA libraries to be run without multiplexing with index barcodes. The surface of the flow-cell is coated with adaptors with complementary sequences to the P5 and P7 adaptors added in the library preparation above. The library material is denatured and added to the flow cell as single-stranded DNA. After hybridizing with surface oligonucleotides the template strand is copied (3). The product is denatured again and the newly synthesized strand stays bound to the surface of the flow-cell while the other material is washed out (4). In the next step the template can bind a nearby oligonucleotide on the surface and hybridize to this, allowing a further strand synthesis to occur. In this way initialization of the solid phase cluster PCR is achieved. A bridging PCR as shown above results in the required clonal amplification of the sequence template (5).

On most Illumina sequencing systems the cluster formation is randomly distributed over the flow-cell and the density, be it over clustered and therefore unreadable or under clustered producing a poor yield, is highly dependent on the concentration of the DNA library added. The size selection of the fragments is also essential to ensure that the PCR products stay within the formed clusters, otherwise the sequence data cannot be unambiguously read. A general rule of thumb would be a cluster density of 610–678 K mm^{-2} for the HiSeq 2000 platform although other sequencers of the Illumina family require different densities. Illumina has produced patterned flow-cells for their later sequencing platforms HiSeq x10 and HiSeq 4000 that allow higher loading of the flow-cells by restricting cluster formation to fixed well positions on the flow-cell.

Once prepared the flow-cell is loaded onto the sequencer. In the first sequencing step all DNA molecules with polymerization capable ends are blocked to stop unspecific primer extension. The sequencing process is then initialized by adding a mix of sequencing primers, polymerases, and four different fluorescent markers, reversible protected dNTPs. Owing to the extension-protected nucleotide each cluster has just one base incorporated complementary to the template strand (6, Figure 30.22). Finally, at the end of each sequencing cycle the flow-cell is scanned, producing four pictures, one for each of the fluorophore laser excitement frequencies representing one for each base. Afterwards the fluorophores are cleaved off and the end of the sequence strands reactivated for further polymerization. The sequence is called based on the positional information relating to the cluster and the sequence of base specific fluorescence (7).

The maximum output from this system is governed by the flow-cell surface available to cluster and the read length, which is itself dependent on the speed at which the fluorescent sequencing cycles can be performed. Illumina has various different models that offer higher throughput rates or longer read lengths. Currently, highest throughput occurs on a HiSeq 4000 with 2× 150 base reads giving around 1.5 terabase of data in 3 days. Longer read lengths but much less data is produced by the MiSeq, giving read lengths of over 2 × 300 bases.

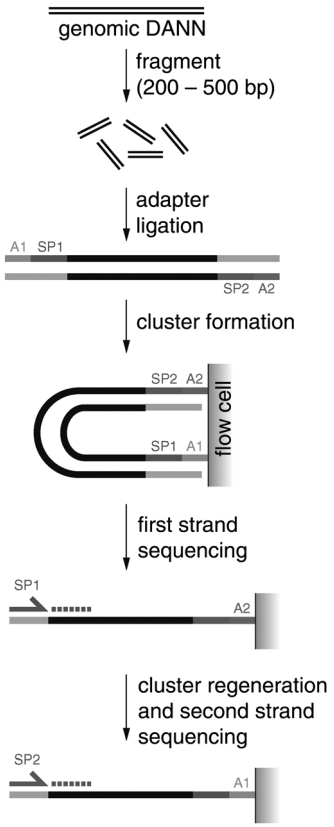


Figure 30.23 Pair-end sequencing: adapters A1 and A2 allow bridge amplification from either end of the strand and sequencing primers SP1 and SP2 allow sequencing to be started from each end sequentially.

Adaptations to Library Preparation and Sequencing The amount and properties of the sequence reads produced by sequencing such as read length and read number are governed by the capabilities of the sequencing instrument and the chemistry it uses. This often means that sometimes the read sequence is too short to be used in certain steps such as spanning over a repeat region, or that there are not enough reads to cover a whole genome. Certain adaptations to the library preparation and sequencing techniques have been developed to deal with this. The examples given below are taken from the Illumina platform but can be applied to other platforms as well.

Paired-End-Sequencing When the template is sequenced from both ends (Figure 30.23), the reactive group mentioned above is cleaved after the first end is read. The reverse strand can then cluster via bridging PCR and then be sequenced in the same way the first read is sequenced. The advantages of paired end sequencing are (i) the extra coverage of a second read and (ii) the positional information provided by having two anchor reads of known distance apart. Paired end reads are often used in genomic sequencing, allowing structural variations in the sequenced genome to be identified, and assist genomic assembly considerably (Figure 30.24).

Mate-Pair Library Creation This refers to the generation of a specific type of library to obtain sequence anchors spanning a wider sequence distance than can be handled with normal paired end reads. Hence the libraries are formed with a large inner distance between the sequence fragment end, circularized, and the shorter outer fragment is used to form paired end reads. Here the goal is to obtain positional information from separated anchor reads and is often

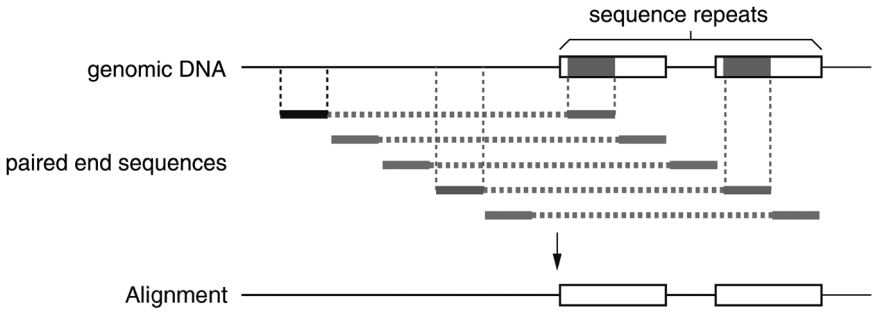


Figure 30.24 Alignment of paired-end-sequences allows both ends of a sequenced fragment to be aligned. This allows unique alignment of one of the pairs outside a repeat when the other end aligns in a repeated region, allowing repetitive regions to be unambiguously aligned and making assembly of repeated sequences easier.Source: adapted after Illumina, Inc.

used in scaffolding of *de novo* assemblies (Figure 30.25). For this type of library the first step is to fragment the DNA to 5 kb or greater instead of 200–500 bp. This gives a large spanning distance between the read that can be used to span over repeated regions in the genome. As fragments of this length cannot be used for clustering on the flow-cell the ends are biotinylated and the fragments circularized. The circularized DNA is once more fragmented (400–600 bp) and the biotinylated region captured. The captured fragments are then sequenced as normal in the paired end fashion.

Indexed Libraries As the number of sequence reads per sequencing run increases, the question arises: How deep does one have to sequence a particular sample? If it is the case that a desired sequence can be covered with less depth than would be delivered by one lane or run of the sequencing several samples can be run together with sequence barcodes included during library preparation (Figure 30.26). In the case of Illumina sequencing a library is created as described above. After adaptors are ligated that already carry the sequencing read primer 1 the fragments are PCR amplified. Here the index barcode is introduced to the library fragments. One of the PCR primers has the index sequence and the sequencing primer for the second read. A second primer contains the P5 structure for attachment to the flow-cell surface. A third primer carries the P7 attachment section and the index sequence. The first set of index barcodes consisted of 12 modified P7-adapter sequences to be used so several different samples could be run in the same flow-cell lane. The index is separately read through an index read and can be up to eight bases long. Now, the use of dual indexing barcodes of 16 bases is possible. Other methods involve the integration of the index after the standard sequencing primers. However, with this method there is the disadvantage that there is a loss in sample sequence length as the index is read at the start of the first sequencing run.

Target (Exome) Enrichment When it is uneconomical to sequence a whole genome or genomes of individuals or if only a particular genomic region is required to be sequenced, target enrichment can be used to extract only the desired regions of the genome for sequencing. The most frequent example of this is exome enrichment or exome capture. This method reduces the sequencing to the coding regions of the genome and allows comparison of the exome sequences from a range of individuals. Illumina technology is also used as an example here (Figure 30.27). After denaturing, the fragments of the DNA library are hybridized to a capture library of biotinylated probes and then bound to magnetic streptavidin-coated beads and eluted. The enrichment process is repeated and the eluted capture product is amplified before sequencing. The produced inserts cover up to 460 bases around the center of the probe and therefore can also include exon flanking regions. The methods of other companies are based on similar principles but often use other probe sequences to capture a different subset of the genome.

Semiconductor Sequencing (Ion Torrent) This method is also based on DNA sequencing by synthesis. The material is also handled similarly to the above processes (creation of the library and clonal amplification). The technology here, however, is based on a different form of

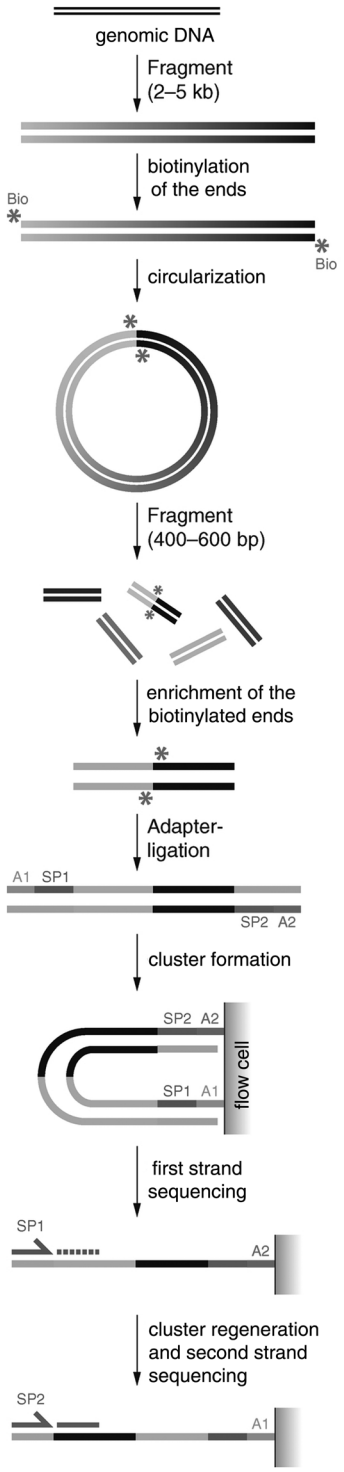


Figure 30.25 Mate-pair sequencing uses biotinylation, circularization, a second fragmentation, and then ligation of paired end sequencing adaptors to the ends of the original molecule. This allows the sequencing of the ends of a fragment, which can be in the range of 5 kb long, giving spatial information about the genome. Source: adapted after Illumina, Inc.

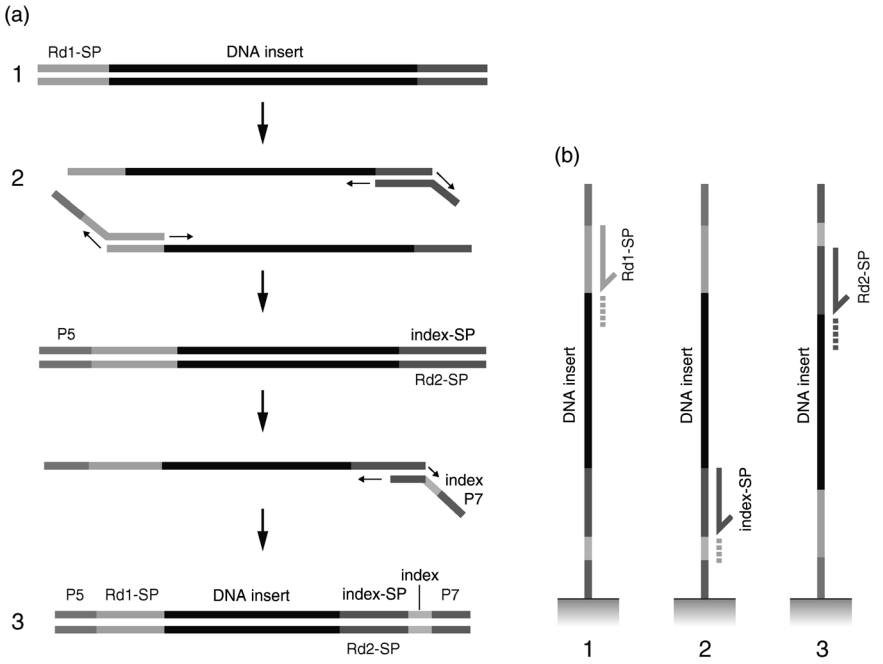


Figure 30.26 Indexed libraries: (a) creation of the library. 1, Rd1-SP-adapter ligation; 2, PCR to add the P5 and index-SP and Rd2-SP and P7 sequences with index barcode sequence; 3, structure of the sequencing template. (b) During sequencing three independent read cycles are carried out: 1, read 1; 2, index read after the removal of the first strand; 3, read 2 of the library molecule. The index read is a separate read and in this way does not reduce the length of either read 1 or 2.

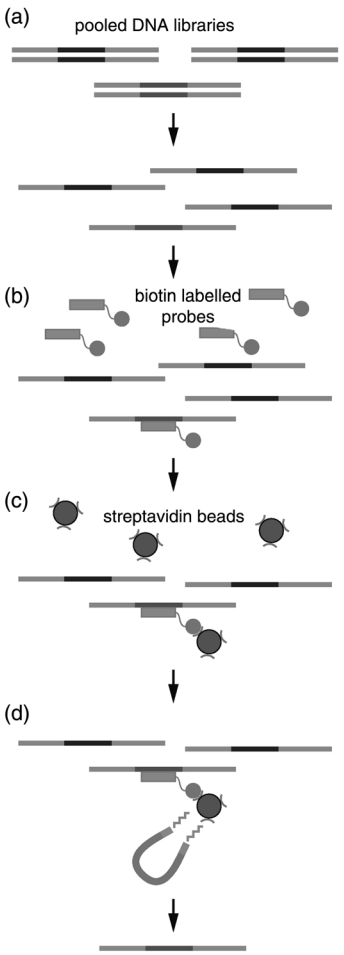


Figure 30.27 Exome enrichment for Illumina systems. (a) Denaturing of the dsDNA library. (b) Hybridization of biotinylated capture probes to the target regions. (c) Enrichment of the target regions using streptavidin beads. (d) Elution of the target fragments from the capture beads. Source: adapted after Illumina, Inc.

detection. With every DNA polymerase catalyzed extension of a nucleotide chain a hydrogen ion (proton) is released that leads to a change in the pH of the reaction volume. In other systems the pH changes are buffered by the reaction mixture, in this method they are used to detect an incorporation of a base. The design of the system transfers a great deal of the sequencing device onto a disposable silicon based sequencing chip. The sequencing reaction takes place on this microfluidic chip in which several cavities are loaded with DNA bound beads. Each nucleotide base is flooded in cycles into the reaction chambers in series individually so that a change of pH measured in the presence of a particular nucleotide indicates which base was incorporated. The charge changes on the sensory surface at the base of the chamber are measured as a voltage change and are used to produce a fluorogram of the sequence. The level of voltage change is used to measure how many bases were incorporated (Figure 30.28).

Applications of Sequencing Technologies NGS sequencing has proven exceptionally useful in scientific research. Although the sequence reads are short the mass of them together provides a large amount of data that can be put to good use.

Genomic Sequencing and Resequencing Sequencing the first human genome was an exceptionally expensive and lengthy project – estimated cost $3 billion. NGS technologies now allow resequencing of a human genome for under $1000 and in under a week. With the availability of cheap mass sequencing the genomes of organisms that would not normally be sequenced can be constructed although several issues exist using the short read data.

If no reference genome is available one has to be assembled *de novo*. This is a challenging task especially for short read technologies where the read length is often shorter than the length of repeated sequences that are widely spread through eukaryotic genomes. Several programs using a range of techniques such as de Bruijn or String figures have been published to deal with the task of constructing larger contiguous sequences, contigs, from short reads. One way to try and extend the capacity of short reads over the length of short repeats is by using paired and mate-pair sequencing strategies.

Despite the best efforts of biologists and computer scientists the repeat nature of eukaryotic genomes has led to several newly genome assemblies that are unfortunately made up of fragmented contigs and would benefit from the promise of the third generation of sequencing technologies presented below.

In cases such as humans where a good reference genome exists it is possible to produce a consensus by aligning reads to the reference. This is widely used in human genome resequencing where variation from the reference is being studied. Single base and small deletions can be detected by modern alignment programs whereas larger structural variants, inversions, deletions, or chromosomal rearrangements may be detected using paired or mate-pair information.

In some cases it is not required to sequence the whole genome of an individual but only a reduced portion such as only the coding sequences for the genes, the exome. This approach has been heavily exploited in disease screening in humans where it was at one stage still impractical to sequence the whole genome. With the advent of higher throughput from the sequencers many large projects have moved to whole genome sequencing with the advantages of more structural variant and non-coding variant information. However, exome sequencing still provides a good way to sequence the gene coding information from several individuals.

Tag Counting – RNASeq and ChipSeq The production of millions if not billions of small sequence reads lends itself very well to various semi-quantitative and qualitative techniques that use different ways to isolate nucleic acids with the specific aim of investigating a particular biological question. The field has expanded widely, encompassing 70 techniques (Illumina has published a summary of available methods – ForAllYouSeqMethods.pdf, available online). Two of the earliest and most widely used techniques, RNASeq and ChipSeq, are outlined below.

RNASeq (Mortazavi, A. (2008) *et al.* Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods*, **5**, 621–628) allows the investigation of gene expression by measuring RNA levels in a cell or tissue. Total RNA is isolated from the tissue and enriched for coding sequences by either capturing poly-A tail containing molecules or by depletion of ribosomal RNA with probes that perform specific sequestering. The RNA is then converted into a single stranded cDNA library using random hexamer priming, following which the second strand of the cDNA is synthesized, creating a library of cDNA molecules that can be further processed in library preparation.

Once the library is prepared it is sequenced as normal on the sequencer. The short read tags are then aligned to a genome and the alignment positions mapped back to the relevant gene annotation. The number of aligned counts per gene, transcript, or exon is taken as a measure of gene expression, with higher tag counts indicating higher gene expression. Paired end sequencing or aligning with an aligner capable of splitting the short read alignments up so that they span exon junctions are used when investigating alternative transcript usage in gene expression.

ChipSeq (chromatin-immuno-precipitation sequencing) (Robertson, G. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4,** 651–657) is used to investigate the binding sequences of DNA binding proteins such as transcription factors.

DNA and bound proteins (chromatin) are reversibly crosslinked using formaldehyde. The chromatin is sheared and enriched by immunoprecipitation of the DNA binding the protein using an antibody specific for the protein in question. After enrichment the crosslinking is reversed and the DNA is processed into a sequencing library and sequenced. The short reads are aligned against a genome and the number of tags aligning above background and the shape of the peak of aligned reads is used to ascertain the binding position of the protein under investigation. In the case of transcription factors the question is often: Which gene is downstream of the binding site?

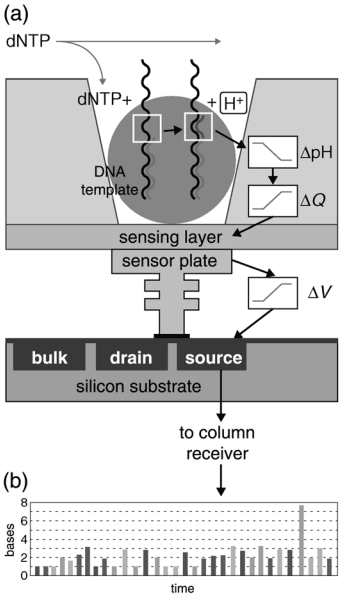


Figure 30.28 Ion torrent: (a) Single reaction chamber of a sequencing chip with a single bead bound DNA template, the sensor, and electronics. With the addition of dNTPs, protons (H^+) are set free that change the pH of the chamber. These changes are measured by the chamber's sensor. (b) Variation of bases with time. The peak heights of the figure represent the number of bases detected. The bases incorporated are identified by the step in the synthesis and the dNTP added. Source: adapted after Ion Torrent Systems, Inc. (now Thermo Fisher Scientific GENEART GmbH).

30.2.2 Single Molecule Sequencing

The above established sequencing methods require an amplification, either by classic cloning in the "older" processes or by clonal amplification in the newer massively parallel sequencing techniques. The availability of new more sensitive detectors and methods based on other

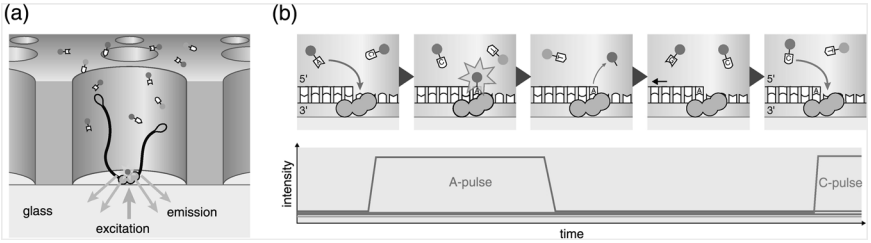


Figure 30.29 PacBio RS single molecule real time DNA sequencing. (a) Single ZMW with a polymerase bound to the lower surface of the chamber. The illumination from underneath only affects the lower region of the chamber. (b) Diagram of the sequencing process (incorporation, fluorescence emission, cleavage of the fluorophore). Source: Eid, J. *et al*. (2009) *Science*, **323**, 133–138. With permission, Copyright © 2009, American Association for the Advancement of Science.

principles brings the possibility of measuring single molecules within reach. This has brought on the concept of third-generation sequencing techniques focused on single molecules.

Single Molecule Real Time (SMRT) DNA Sequencing (Pacific Biosciences) This technique sees a single DNA molecule being sequenced in a very small chamber called a zero mode waveguide (ZMW) in a sequencing cell. There are thousands of ZMWs in the chip and thereby thousands of molecules are sequenced in parallel and also in real time. The special thing about the ZMW is that it is so narrow that light is trapped at its base when emitted from sets of fluorescently labeled nucleotides due to nucleotide incorporation:

1. The starting material is fragmented and end repaired. A hairpin adaptor is ligated on each end resulting in a closed circular molecule. The hairpin structure contains a sequencing primer complementary sequence. The final products are size selected and cleaned.
2. The sequencer primer is hybridized onto the hairpin structure. After which Phi-29 polymerase is added and an initiation complex formed that can be loaded onto the chip and sequenced.
3. Once immobilized in the ZMW the sequence of the DNA template is measured by the fluorescence of incorporated molecules. Each base has a different fluorescent label and when incorporated leads to a large burst of light emitted that is trapped for some time in the ZMW, giving a clearly measurable signal. This clear incorporated signal is much stronger than that of unincorporated labeled nucleotides that move in and out of the volume at the bottom of the ZMW quickly. The terminal phosphate carrying the base specific fluorophore is removed with the addition of the next labeled nucleotide and the subsequent cleaving of the phosphate. This results in a real time film of bases being incorporated, light being emitted, next base being incorporated, and the next burst of light being emitted (Figure 30.29).

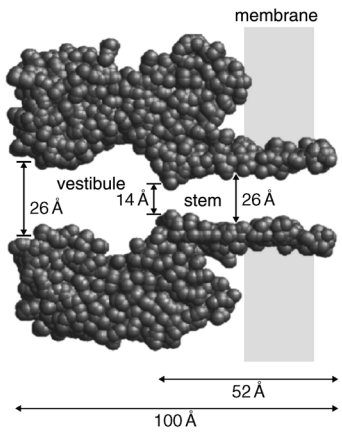


Figure 30.30 Native hemolysin pore incorporated in a membrane. Source: Zwolak, M. and Ventra, M.D. (2008) *Rev. Modern Phys.*, **80**, 141–165. With permission, Copyright © 2008, American Physical Society.

Nanopore Sequencing A nanopore is a very small pore that will allow ions to flow through when a voltage is applied to it in a conducting fluid. The pattern of current generated by the flow of ions is characteristic based on the pore shape and changes also in a characteristic fashion when DNA is threaded through the pore. By threading DNA through a pore the sequence of nucleotides is called by the characteristic changes observed due to different nucleotides as they pass through (Figure 30.30).

This technique promises several potential advantages over other previous techniques in that it can measure a DNA molecule of any length without being restricted to a set of run cycles and that it measures a single molecule without need for amplification. The difficulties with the technique lie in the small changes in current, which are difficult to measure accurately enough to reliably call the base sequence.

Despite the practical difficulties in measuring DNA flowing through nanopores some companies have attempted to bring a nanopore sequencer to market. Oxford Nanopore Technologies is the nearest to commercially launching a product at the time of writing. Using protein nanopores placed in an artificial non-conducting chip surface, DNA is threaded through the pore by a progressive engine attached to one end of the DNA during library preparation, while the other end receives a hairpin sequence. As the DNA passes through the pore at its narrowest point the sequence causes a change in the current flowing and this is measured by the chip's detectors under each pore. When one strand of the DNA strand has been sequenced the hairpin passes through the pore and the second strand is also sequenced, allowing the information from both strands to be used in base calling.

Further Reading

Ansorge, W., Voss H., and Zimmermann, J. (1996) *DNA Sequencing Strategies*, John Wiley & Sons, Inc., New York.

Bentley, D.R. *et al*. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Blazej, R.G. *et al*. (2007) Inline injection microdevice for attomole-scale Sanger DNA sequencing. *Anal. Chem.*, **79**, 4499–4506.

Branton, D. *et al.* (2008) Nanopore sequencing. *Nat. Biotechnol.*, **26**, 1146–1153.

Brenner, S. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18**, 630–634.

Clarke, J. *et al.* (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.*, **4**, 265–270.

Craig, D.W., Pearson, J.V., Szelinger, S. *et al.* (2008) Identification of genetic variants using bar-coded multiplex sequencing. *Nat. Methods*, **5**, 887–893.

Deamer, D. (2010) Nanopore analysis of nucleic acids bound to exonucleases and polymerases. *Annu. Rev. Biophys.*, **39**, 79–90

Drmanac, R. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays, *Science*, **327**, 78–81.

Fuller, C. *et al*. (2010) The challenges of sequencing by synthesis. *Nat. Biotechnol.*, **27**, 1013–1023.

Hodges, E. *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.*, **39**, 1522–1527.

Horner, D.S., Pavesi, G., Castrignano, T. *et al.* (2009) Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings Bioinformatics*, **11**, 181–197.

Korlach, J. *et al.* (2008) Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 1176–1181.

Levene, M.J. *et al.* (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, **299**, 682–686.

Mardis, E.R. (2003) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Human Genet.*, **9**, 387–402.

Margulies, M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Maxam, A. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 560–564.

Niedringhaus, T.P. *et al*. (2011) Landscape of next-generation sequencing technologies. *Anal. Chem.*, **83**, 4327–4341.

Pop, M. and Salzberg, S.L. (2008) Bioinformatics challenges of new sequencing technology. *Trends Genet.*, **24**, 142–149.

Ramanathan, A., Huff, E.J., Lamers, C.C. *et al.* (2004) An integrative approach for the optical sequencing of single DNA molecules. *Anal. Biochem.*, **330**, 227–241.

Richardson, P. (2010) Special issue: next generation DNA sequencing. *Genes*, 385–387.

Rothberg, J.M., Leamon, J.H. (2008) The development and impact of 454 sequencing. *Nat. Biotechnol.*, **26**, 1117–1124.

Rothberg, J.M. *et al*. (2008) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–352.

Sanger, F., Nicklen, S., Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, **74**, 5463–5467.

Schadt, E.E. *et al.* (2010) A window into third-generation sequencing. *Hum. Mol. Genet.*, **19**, R227–R240.

Schuster, S.C. *et al.* (2008) Method of the year, next-generation DNA sequencing. Functional genomics and medical applications. *Nat. Methods*, **5**, 11–21.

Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.

Stoddart, D. *et al*. (2009) Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 7702–7707.

Stoddart, D. *et al.* (2010) Nucleobase recognition in ssDNA at the central constriction of the alpha-hemolysin pore. *Nano Lett.*, **10**, 3633–3637.

Sultan, M. *et al*. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.

Tabor, S. and Richardson, C.C. (1990) DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *J. Biol. Chem.*, **265**, 8322–8328.

Trepagnier, E.H. *et al.* (2007) Controlling DNA capture and propagation through artificial nanopores. *Nano Lett.*, **7**, 2824–2830.

Venter, J.C., Adams, M.D., Meyers, E.W. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Wang, H. and Branton, D. (2001) Nanopores with a spark for single-molecule detection. *Nat. Biotechnol.*, **19**, 622–623.

Analysis of Epigenetic Modifications

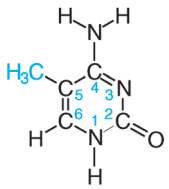
31

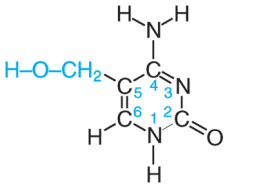
Reinhard Dammann

Justus Liebig-University of Gießen, Institute of Genetics, Department of Biology and Chemistry, Heinrich-Buff-Ring 58-62, 35392 Gießen, Germany

With the complete sequencing of the human genome the number of genes that are involved in the complex interaction of the cellular development is now assessable. Basically, each genome consists of four bases: adenine, thymine, cytosine, and guanine, which, however, can be modified covalently. Since these modifications are inherited after DNA replication, the coded information of the genome changes considerably. The most important DNA modifications are methylation of cytosine at the C5-postion to 5-methylcytosine (^{5m}C) and the methylation of adenine at the N6-position to N6-methyladenine (^{N6m}A) (Figure 31.1). Methylation of adenine is mainly found in prokaryotes (Dam-methylation, (G^{N6m}ATC) and serves as a protective mechanism of its own DNA against sequence specific restriction enzymes.

Cytosine methylation is frequently found in bacteria (Dcm-methylation, C^{5m}CWGG) and is also detected as a modification in plants, invertebrates, and vertebrates (i.e., CpG-methylation, ^{5m}CG). In mammals cytosines are mainly methylated, when the base is followed by guanine, which is designated as dinucleotide ^{5m}CpG. This methylation is achieved *in vivo* by DNA methyltransferases (DNMTs). Interestingly, an additional ^{5m}CpA- and ^{5hm}C-methylation has been found in human stem cells. In addition, in the brain, 5-hydroxymethylcytosine (^{5hm}C) has been detected and results from the oxidation of ^{5m}C by TET enzymes. TET further oxidizes ^{5hm}C to 5-formylcytosine (^{5f}C) and 5-carboxylcytosine (^{5ca}C), which can be converted into C by base excision repair. The function of ^{5hm}C is not yet completely understood and may either represent an intermediate modification in the DNA demethylation pathway or a specific epigenetic mark. In human somatic cells ^{5m}C base makes up only 1% of all DNA bases, but 70–80% of the CpG are methylated. In the human genome the dinucleotide CpG is underrepresented, but is often found in GC-rich sequences, so-called CpG islands. Nearly 60% of all human genes harbor a CpG island in their promoter region. Normally, these CpG-island promoters are unmethylated. Methylation of CpG in the promoter region has an important influence on the regulation of gene expression and leads to epigenetic inactivation of the affected gene. Epigenetic control plays an important role in inheritance of gene activity; since DNA methylation modifies the information of the genome without changing the primary sequence pattern of the DNA. Cytosine methylation directly influences the gene activity by the binding of regulatory proteins (e.g., methyl binding domain proteins) to the methylated sequence and indirectly influences gene expression by inactivation of the chromatin structure and through altered histone modifications. Thus, the fifth base ^{5m}C acts as a reversible epigenetic switch and is essentially involved in the inheritance of the gene activity. Since hypermethylation of regulatory sequences results in inactivation of gene expression, these epigenetic changes are considered to be an important mechanism in the inactivation of tumor suppressor genes. DNA-methylation not only plays a fundamental role in carcinogenesis but also in cellular development and aging. Furthermore, DNA-methylation determines allele specific expression of paternal and maternal inherited genes, a mechanism that is termed imprinting. DNA methylation is also involved in the processes of dosage compensation by X-chromosome-inactivation. Moreover chromatin structure and modifications of nucleosomes are also important for the epigenetic gene





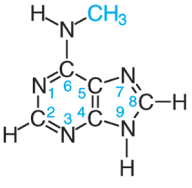


Figure 31.1 5-Methylcytosine, 5-hydroxymethylcytosine, and N6-methyladenine.

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

regulation. While active genes display an open chromatin structure (euchromatin) with acetylated histones, silenced genes show a closed chromatin (heterochromatin) with deacetylated histone. Additionally, nucleosomes are altered by methylation, phosphorylation, and other modifications of histones. This chapter describes several methods that can be used to analyze epigenetic modifications (DNA methylation and other chromatin changes).

31.1 Overview of the Methods to Detect DNA-Modifications

There are six main techniques for analyzing DNA-methylation:

- Chemical modifications of the unmethylated bases with:
  1. bisulfite.
- Protein specific analyses of DNA-sequences with:
  2. methylation sensitive restriction enzymes;
  3. ^{5m}C-binding domain (MBD) proteins;
  4. ^{5m}C, ^{5hm}C, ^{5f}C or ^{5ca}C -specific antibodies.
- Analysis of the configuration of the bases of the complete DNA with:
  5. DNA hydrolysis;
  6. nearest neighbor-analysis (Table 31.1).

The different reactivities of modified cytosine and cytosine for bisulfite induced deamination to uracil is used to determine the methylation status of the DNA. With protein specific methods the different activity of restriction enzymes or binding of ^{m}C-binding proteins (MBD or antibodies) is used to analyze the methylation status of the DNA. Certain methylation sensitive restriction nucleases do not cut their recognition site if the DNA is methylated, whereas others only cut methylated sites in the DNA or are insensitive against methylation. These enzymes allow analysis of the methylation status of the respective restriction cutting sites. ^{5m}C-Binding domain proteins (MBDs) and ^{m}C-antibodies (^{5m}C or ^{5hm}C) are utilized to precipitate the modified DNA and to quantify its methylation status by real time PCR. Moreover, antibodies that recognize ^{5f}C or ^{5ca}C are available. A further possibility is in analyzing the composition of the DNA bases. The genomic DNA is completely hydrolyzed and different modifications of a

Real Time PCR, (quantitative) PCR, Section 29.2.5

Table 31.1 Overview of important methods for analysis of the DNA-modifications.

Method	Concept	Identification	Scope
Bisulfite modification	Chemical resistance of ^{5m}C and ^{5hm}C to deamination to uracil by bisulfite (^{5f}C, ^{5ca}C, and C are deaminated to U)	– Sequencing – restriction analysis (COBRA) – methylation specific PCR (MSP) – TET assisted bisulfite sequencing (TAB-Seq)	All modified Cs in a DNA fragment can be analyzed as well in the upper as in the lower strand
Methylation sensitive restriction enzymes	Different accessibility of methylated DNA for restriction enzymes	– Southern blotting – qPCR	Only DNA-modifications within restriction site can be analyzed
^{5m}C-Binding domain (MBD) protein-specific analyses (i.e., MIRA)	Precipitation of DNA or chromatin with MBD proteins	– Pulldown and sequencing – qPCR – microarray – immunofluorescence	The methylation status of a specific DNA fragment or region can be analyzed
Antibody specific for modified DNA (i.e., MeDIP)	Precipitation of DNA with ^{5m}C-, ^{5hm}C-, ^{5f}C-, or ^{5ca}C-antibodies	– Immunoprecipitation and sequencing – qPCR – microarray – immunofluorescence	The methylation status of a specific DNA fragment or region can be analyzed
DNA hydrolysis	Complete analysis of the different base modifications	– HPLC – chromatography – mass spectrometry	Different modifications of the genomic DNA can be analyzed
Nearest neighbor-analysis	Analysis of the different DNA modifications in connection to the 3′-base	– HPLC – chromatography – mass spectrometry	The amount of distinct modifications can be analyzed

single base are analyzed. With the nearest neighbor-analysis dinucleotides are labeled and subsequently their composition is dissected. However, DNA hydrolysis and the nearest neighbor-analysis are not able to reveal the exact sequence context of the modified bases. These methods are explained next in more detail.

31.2 Methylation Analysis with the Bisulfite Method

The easiest and most effective way to analyze DNA methylation is by the bisulfite technique. This method has a very high resolution and it is possible to analyze the methylation status of the whole DNA-population at a specific sequence or to dissect the methylation pattern of single DNA fragments. This technique was developed in 1974, but it was only in 1992 that it became popular after further development by Frommer and coworkers. Meanwhile it has come into broad usage because of its high resolution and reliability. The principle of this method is the reaction of bisulfite (HSO_3^-) with DNA, which converts cytosine (^{5f}C, ^{5ca}C, and C) into uracil. The C6 position of an accessible cytosine is sulfonated by a high bisulfite concentration (3.0 M) and acidic conditions (pH 5.0). In this process the amino group at the C4 position is hydrolyzed and uracil is generated (Figure 31.2). The particularity of this reaction is that methylated cytosines (^{5m}C and ^{5hm}C) are not converted and remain as cytosines. Thus, unmethylated cytosines (deaminated to U) and methylcytosines (remain C) can be distinguished. (Figure 31.3). However, this method cannot distinguish between ^{5m}C and ^{5hm}C. In a PCR reaction the bisulfite-treated DNA is amplified and analyzed by using primers, which are complementary to the deaminated DNA sequence. This PCR results in the substitution of uracil by thymine (Figure 31.2). Through the PCR amplification the bisulfite method is highly sensitive and only a small amount of genomic DNA (50 ng) is needed. It is also possible to analyze the DNA methylation in less than 100 cells. Before the bisulfite treatment it can be of advantage if the cells or the DNA are embedded in agarose so that the loss of DNA can be minimized. A problematic issue of the bisulfite treatment is the incomplete conversion of Cs into Ts. The partial denaturation of the DNA during the bisulfite treatment may lead to incomplete deamination of unmethylated Cs to Ts and this could lead to misinterpretation of methylated Cs. To overcome this problem, different modifications of the bisulfite method have been established. One possibility is to digest the DNA in short fragments with a restriction enzyme before its denaturation. (However, no restriction site should be in the investigated DNA region.) During the bisulfite treatment, by a repetitive denaturing in a thermocycler, the deamination of the DNA can also be improved. However, by too intensive treatment, the DNA could be degraded into small fragments. It is relatively simple to verify the complete bisulfite conversion of the DNA by PCR amplification and sequencing. The presence of methylated C in a non-CpG context is mostly likely an artifact of an incomplete bisulfite reaction and this should be verified with an alternative method. In practice the bisulfite technique has proved to be a very efficient and reliable procedure to analyze DNA methylation.

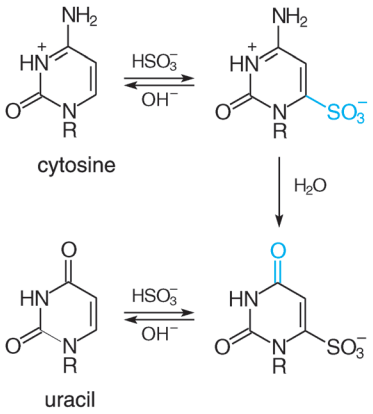


Figure 31.2 Sodium bisulfite catalyzes the deamination of unmethylated cytosine or to uracil.

31.2.1 Amplification and Sequencing of Bisulfite-Treated DNA

The bisulfite method allows the gene specific analysis of methylation levels of a cell population and of the methylation pattern of single DNA molecules. This depends on whether PCR products are directly sequenced (i.e., pyrosequencing) or if they are first subcloned and then sequenced. Since the two DNA strands of the DNA are no longer complementary after bisulfite conversion, it is possible to investigate strand specific methylation with separate primer pairs (Figure 31.3). The PCR amplification leads to a conversion of U (unmethylated C) into T, and from methylated C into C. On the complementary strand G (opposite to an originally unmethylated C) is then converted into A. It is very easy to mimic the bisulfite conversion *in silico* with a word processing program to generate the deaminated sequence necessary to design primers for the amplification of the bisulfite-converted DNA.

Sequencing by Synthesis, Classic Pyrosequencing, Section 30.2.1

The following aspects should be considered when designing the primer:

1. To exclude the amplification of not bisulfite-modified DNA, primer pairs should include some deaminated C (T in forward-primer and respective A in the reverse primer).

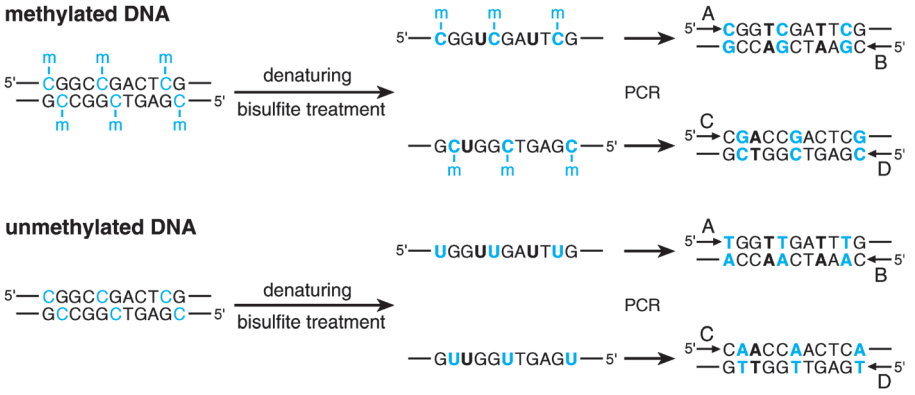


Figure 31.3 Principle of the methylation analysis with the bisulfite method. DNA is denatured and treated with bisulfite. With this method the methylated cytosines (^{5m}C and ^{5hm}C) are conserved while the unmethylated cytosines (C, ^{5f}C and ^{5ca}C) are deaminated to uracil (U) and appear as thymine (T) after PCR amplification. Note that after the bisulfite treatment the DNA strands are not complementary and can be amplified with different primer pairs (A, B or C, D).

2. The primer should not include CpGs in their original DNA sequence, to avoid a specific amplification of methylated or unmethylated DNA (see also methylation specific PCR). If this is unavoidable, it is possible to insert Y (pyrimidine: C or T) instead of C and in the complementary strand an R (purine: G or A) instead of G.
3. Since primers do not contain Cs (in the complementary strand no Gs) their annealing temperature is often low and therefore primers with a length of 25–30 nt should be used.
4. For pyrosequencing a biotinylated primer and a sequence primer are necessary.
5. The PCR product should not be longer than 500 bp since longer DNA fragments will be amplified at low rate. This is caused by the fact that the bisulfite treatment degrades DNA and long intact DNA molecules are not available for the amplification.
6. The methylation pattern of individually complementary DNA molecules can be analyzed. Methylation in the context of a double strand DNA can be analyzed by the ligation of a *hairpin linker* prior to the bisulfite treatment.

Semi-nested or Nested PCR, Section 29.3.1

Chemical Cleavage According to Maxam Gilbert, Section 30.1.3

Normally, 50 ng bisulfite-treated DNA is fully sufficient for a PCR reaction. If only a small amount of DNA is available, the sensitivity of the detection can be increased by a semi-nested or nested PCR. The DNA methylation is identified either by direct sequencing of PCR products or by sequencing of many individual DNA molecules after cloning in a vector system. One advantage of the bisulfite method is that the DNA methylation can be detected by conventional sequencing (didesoxy or Maxam–Gilbert method) or pyrosequencing (Figure 31.4). Methylated C as well as unmethylated C can be detected with the DNA sequencing method: ^{m}C occurs as C and unmethylated C as T (Figure 31.4).

Hence, the methylation of all CpG in a DNA fragment can be analyzed. From the pyrosequencing additional quantitative conclusions can be drawn about the level of methylation of the PCR products and the presence of incomplete conversions of unmethylated cytosines (see bisulfite control in Figure 31.4). Alternatively, sequencing from short DNA fragments can be accomplished by mass spectrometry.

During the PCR reaction a preferential amplification (bias) of unmethylated or methylated DNA may occur. In samples representing a certain mixture of methylated DNA and unmethylated DNA, this PCR bias can be investigated. To generate methylated standards DNA is methylated by a CpG-methylase (e.g., *Sss*I-methylase) before bisulfite conversion and if necessary by cloning of PCR products in *Escherichia coli*.

31.2.2 Restriction Analysis after Bisulfite PCR

In Vitro Restriction Analysis, Section 27.1.4

For further analysis of bisulfite-modified DNA different sensitive detection methods have been developed. One of them is the restriction analysis of PCR products of bisulfite-treated DNA. This method was termed combined bisulfite restriction analysis (CoBRA). The principle is that methylated C remains C after bisulfite treatment in a CpG sequence while unmethylated C is converted into T. If the C is in a palindromic sequence (i.e., 5′-TCGA) the methylation can be investigated with a restriction enzyme. The restriction enzyme *TaqI* cuts the recognition sequence TCGA, so that the "methylated" PCR products are digested (Figures 31.5 and 31.6). On the other hand, if this restriction site is missing and the product is not cut, the PCR product originates from an

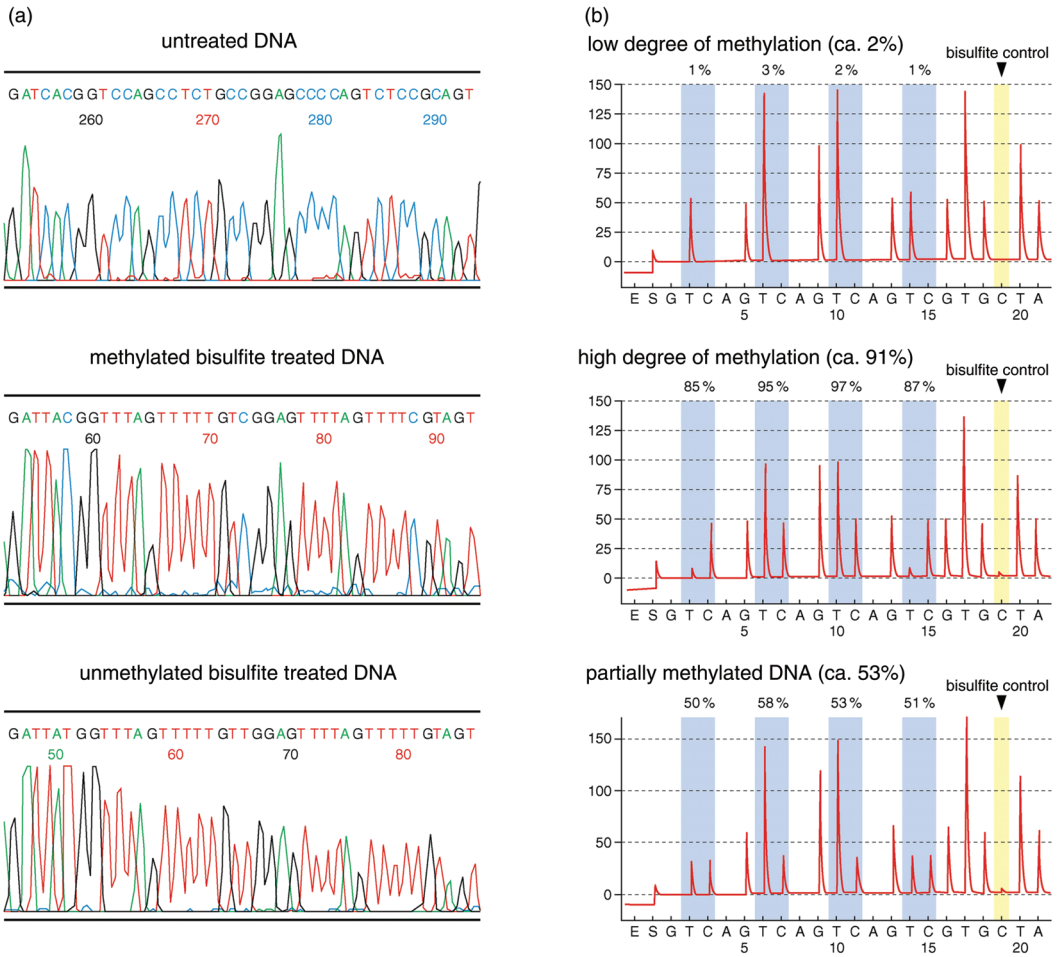


Figure 31.4 Examples of bisulfite methylation analysis after conventional sequencing (a) or pyrosequencing (b) of PCR products of the RASSF1A-promoter. (a) After bisulfite treatment and PCR amplification all unmethylated cytosines (C) are replaced by thymines (T). Methylated Cs in CpG context are resistant against this conversion. (b) Pyrograms of three sequencing reactions with the sequence YGTTYGGTTYGYGTTTGTTA and different levels of methylation of the PCR products. Double height of the signal indicates the incorporation of two nucleotides.

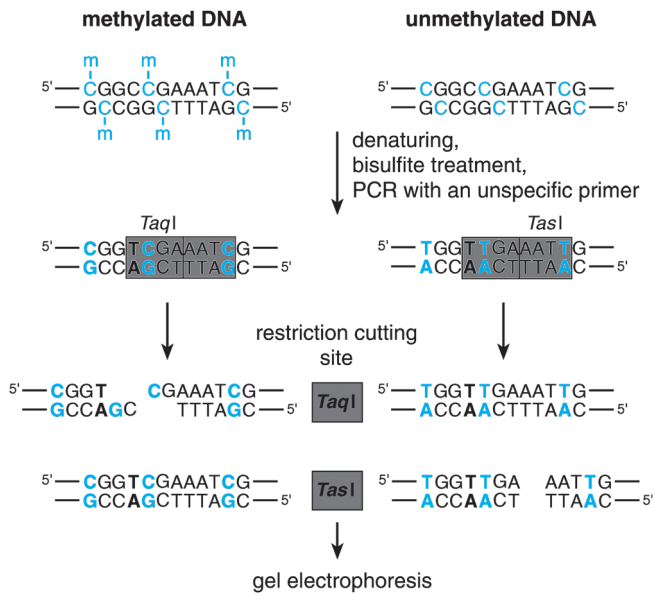


Figure 31.5 Principle of the restriction analysis after bisulfite PCR. The DNA is denatured, treated with bisulfite, and amplified by PCR. While unmethylated cytosines are modified to thymines (T) the methylated cytosines (^{m}C) remain C. For example, in the "methylated" DNA the restriction cutting site for *Taq*I (5′-TCGA) and in the "unmethylated" DNA the recognition site for *Tas*I (5′-AATT) emerges.

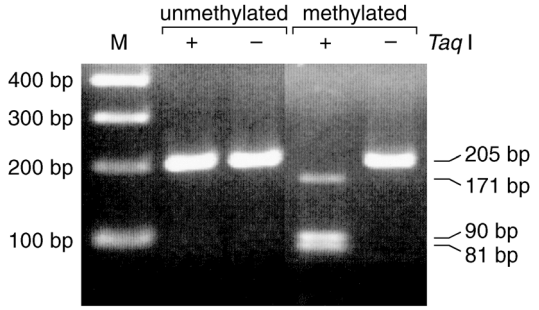


Figure 31.6 Example of a restriction analysis of the RASSF1A-promoter after bisulfite-PCR. The 205 bp long "unmethylated" PCR product is not cut by *Taq*I. The "methylated" PCR product is digested in 171 bp (partial methylated), 90 bp, and 81 bp fragments. (The 34 bp long fragment is not visualized.) A 100 bp marker (M) serves as length standard for the 2% agarose gel.

Table 31.2 Enzymes for the restriction analysis of bisulfite modified DNA of PCR products.

Restriction enzyme	Recognition sequence
For methylated DNA (**C**pG):	
*Taq*I	T/**C**GA
BstUI	**C**G/**C**G
*Mae*II, *Hpy*CH4IV or *Tai*I	A/**C**GT
*Bsi*WI	**C**/GTA**C**G
*Pvu*I	**C**/GAT**C**G
*Cla*I	AT/**C**GAT
*Mlu*I	A/**C**G**C**GT
For unmethylated DNA (**T**pG):	
*Tas*I, *Tsp*509I	/AAT**T**
*Ase*I, *Vsp*I	AT/TAA**T**
*Ssp*I	AAT/AT**T**

unmethylated DNA (Figures 31.5 and 31.6). For COBRA all enzymes with CG in the recognition sequence can be used as diagnostic restriction enzyme (Table 31.2) and also those with only one C at the 3′-end (i.e., *Eco*RI: GAATT**C**). However, the most common enzymes are the four base pair cutter *Taq*I, *Bst*UI, and *Mae*II, since their recognition sites are more abundant (Table 31.2).

Interestingly, this assay can also be used to verify the complete conversion of C into T in the analyzed DNA. Thereby, a new restriction site is created through the bisulfite conversion. For example, a cutting site for *Taq*I (TCGA) is only created from the original sequence 5′-CCGA when the unmethylated 5′-C was modified to T, but the second C is not converted in a methylated CpG context (Figure 31.5).

The same principle can also be used to investigate an unmethylated C in a CpG sequence. If a C at the 3′-end of a putative recognition site is modified to T, a new restriction cutting site is only created in the deaminated "unmethylated" DNA (Figure 31.5). For example, if the sequence AATCG is modified to AATTG, a restriction site for the enzyme *Tas*I (AATT) will be found in a PCR product amplified from unmethylated DNA (Figure 31.5). One main limitation of COBRA is that only the analysis of the DNA methylation at restriction enzyme recognition sites is possible and therefore not all CpG in a DNA molecule can be investigated.

31.2.3 Methylation Specific PCR

In 1996 Herman and coworkers developed methylation specific PCR (MSP) to increase the sensitivity in detecting methylated DNA after bisulfite treatment. MSP analysis is very sensitive and can detect up to 0.1% methylated (or unmethylated) DNA sequences per sample. The MSP method uses different primer pairs for the amplification of methylated and unmethylated DNA after the bisulfite modification (Figure 31.7). These primers are located at specific CpGs and

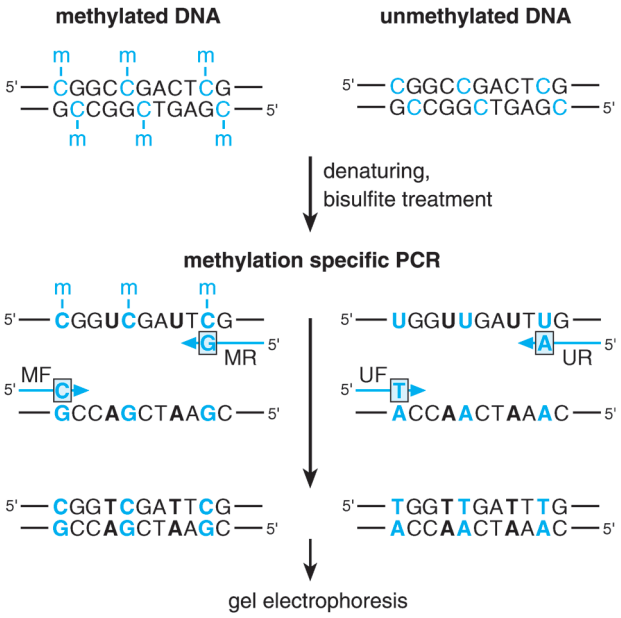


Figure 31.7 Principle of the methylation specific PCR. DNA is denatured and treated with bisulfite. In this process methylated cytosines (^{m}C) are kept while unmethylated Cs are deaminated to uracils (Us). The "methylated" DNA is amplified with methylation specific primer (MF and MR) and the "unmethylated" DNA with unmethylation specific primer (UF and UR).

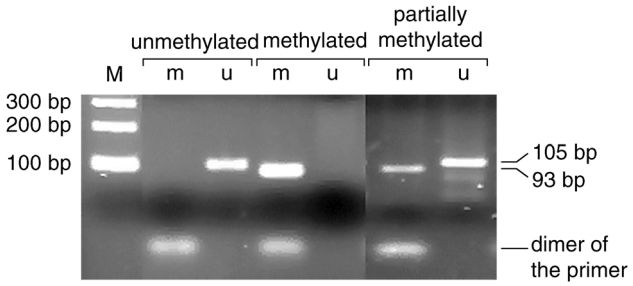


Figure 31.8 Example of a methylation specific PCR of the RASSF1A-promoter. After bisulfite treatment a previously unmethylated DNA is amplified with the unmethylation specific primer pair (u) and a 105 bp PCR product is detectable after gel electrophoresis. On the other hand, a 93 bp PCR product is obtained by amplification of previously methylated DNA with the methylation specific primer pair (m). In a partial methylated bisulfite-treated DNA, PCR products for both primer pairs are detectable. A 100 bp marker (M) served as length standard for the 2% agarose gel.

their amplification rate reveals the methylation status of these Cs. For the amplification of methylated DNA a methylation specific primer pair with Cs in the forward primer and Gs in the reverse primer at the investigated CpG sites is utilized (Figure 31.7). Therefore, these primers only bind and amplify the previously methylated bisulfite modified DNA. In contrast, for the amplification of the unmethylated bisulfite-treated DNA an unmethylated specific primer pair is used, where the C in the forward primer is replaced by T (in the reverse primer G is replaced by A). These primers only bind and amplify the previously unmethylated bisulfite modified DNA.

After gel electrophoresis the methylation status is directly detected by the amount of methylated and unmethylated PCR products (Figure 31.8). To increase the specificity of the amplification rate of "methylated" or "unmethylated" bisulfite-treated DNA by MSP, several aspects should be considered during the primer design. Moreover it is important to utilize DNA controls with known methylation status (i.e. methylated, unmethylated, and unconverted negative controls).

The following aspects should be considered for the primer design:

- The methylation specific forward primer should harbor a C (reverse-primer a G) at the 3′-end.
- The unmethylation specific forward primer should have a T at the 3′-end (reverse-primer an A).
- The primer should have three to four CpG or TpG to ensure a specific amplification of the methylated or unmethylated DNA, respectively.
- To increase the specificity of the primer pairs for bisulfite modified DNA the primer should harbor several Ts for "deaminated Cs" in the forward-primer and As in the reverse-primer.

The advantage of the MSP method is based on its high sensitivity and the simplicity of the assay. MSP has been combined with the real time-detection (real time MSP). The *MethyLight* method uses during the PCR methylation- (or unmethylation-) specific TaqMan probes. Therefore, this technique is utilized to quantify the methylation levels of the analyzed CpGs.

31.3 DNA Analysis with Methylation Specific Restriction Enzymes

Some restriction endonucleases do not cut the DNA when their recognition sites are methylated, while other restriction enzymes are insensitive to such DNA methylation. For a third group of enzymes methylation of their recognition site is necessary for the cutting. These restriction endonucleases are used to identify methylated Cs or methylated As in their recognition sites. Often the enzymes *Hpa*II und *Msp*I are used for the analysis of the methylation status of cytosines at the CpG dinucleotides (Figure 31.9). Both enzymes recognize the sequence 5'-CCGG, but only *Hpa*II is able to cut this sequence, when the second cytosine is unmethylated. In contrast the methylation insensitive isoschizomer *Msp*I is utilized to cut methylated and unmethylated DNA.

*Msp*I cuts the methylated C^{m}CGG-sequence as well as the unmethylated CCGG-sequence. The different methylation specific restriction fragments of *Hpa*II und *Msp*I are analyzed by Southern blot and PCR (Figure 31.9). For Southern blot analysis approximately 10 µg of genomic DNA is necessary. This technique allows a quantitative estimation of the methylated rate at the specific cutting site. Southern blot and hybridization can be conducted by a standard protocol.

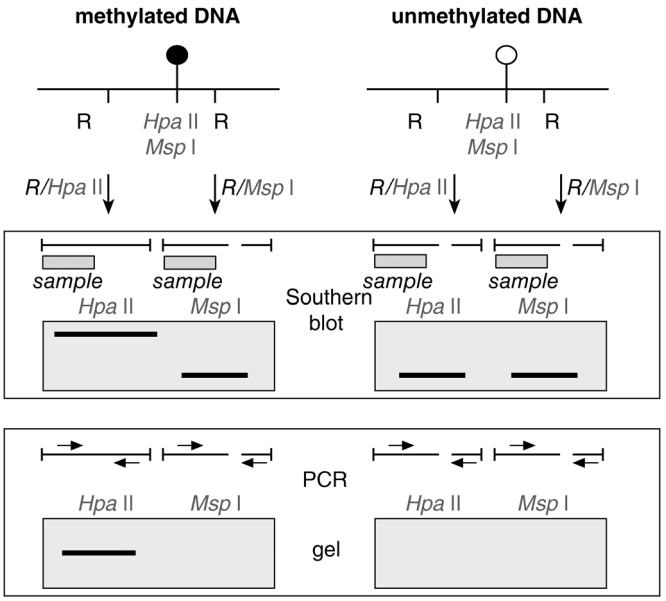


Figure 31.9 DNA-methylation analysis with methylation specific restriction enzymes. The methylation sensitive enzyme *Hpa*II cuts only unmethylated DNA. The methylation specific inhibition is analyzed by Southern blot or PCR and controlled with the insensitive enzyme *Msp*I, which cuts both the methylated and the unmethylated DNA. For the Southern blot analysis the DNA is digested with an additional insensitive restriction enzyme (R).

Isoschizomers are distinct restriction endonucleases with identical recognition sites, which generate similar or different cleavage products.

For the PCR analysis after a methylation specific restriction, little genomic DNA (50 ng) is needed and this allows detection of low levels of methylated DNA. However, this technique is prone to restriction artifacts (e.g., uncomplete digestion, see below) and therefore should be well controlled. For this analysis two primers flanking the restriction site should be designed. PCR products are analyzed by gel electrophoresis and compared to specific controls. Only in the methylated sample is a fragment detected after a restriction digest with the methylation sensitive enzyme *Hpa*II and PCR (Figure 31.9). As control, no fragment should be obtained after digestion with the insensitive *Msp*I and PCR amplification. As further control, the DNA can be digested with a restriction enzyme that cuts outside of the analyzed fragment. After this restriction a PCR product should be detected.

For the methylation analysis of ^{m}CpG a number of methylation sensitive restriction enzymes can be utilized. Several of these enzymes are listed in Table 31.3. With these enzymes not only the methylation of known cutting sites can be investigated, but also novel methylated DNA

Table 31.3 Methylation sensitive enzymes and insensitive isoschizomers for restriction analysis.

Sensitive enzyme (insensitive isoschizomer)	Methylated recognition sequence (isoschizomer)
^{m}CpG-methylation	
*Hpa*II (*Msp*I)	/C^{m}CGG (C/^{m}CGG)
*Bst*UI	^{m}CG/^{m}CG
*Not*I	G^{m}C/GGC^{m}CGC
*Asc*I	GG/^{m}CG^{m}CGCC
*Sma*I (*Xma*I)	CC^{m}C/GGG (C/C^{m}CGGG)
Dcm-methylation – C^{m}CWGG	
*EcoR*II (*Bst*NI)	/C^{m}CWGG (C^{m}C/WGG)
*Sfo*I (*Nar*I)	GGC/GC^{m}CWGG (GG/CGC^{m}CWGG)
*Acc*65I (*Kpn*I)	G/GTAC^{m}CWGG (GG/TAC^{m}CWGG)
Dam-methylation – G^{m}ATC	
*Mbo*I (*Sau3*AI)	/G^{m}ATC (/G^{m}ATC)
*Dpn*II (*Sau3*AI)	/G^{m}ATC (/G^{m}ATC)
*Bcl*I	T/G^{m}ATCA
*Alw*I	GG^{m}ATC(4/5)

regions can be isolated that are resistant to digestion by methylation sensitive restriction enzymes. Methylation-sensitive arbitrarily primed PCR, differential methylation hybridization, and restriction land-mark genome scanning are examples of methods used to identify potentially methylated DNA regions in the genome (see Further Reading).

For the methylation analysis with restriction enzymes a certain caution is necessary: The incomplete digestion of unmethylated DNA with a methylation sensitive enzyme can be mistaken as a partial methylated restriction site. Since the restriction of the genomic DNA can be inhibited by contamination of the sample with cell membranes, carbohydrates, or lipopolysaccharides or by utilizing wrong conditions during the reaction (e.g., salt concentration or pH), the purity of the genomic DNA is essential.

In *Escherichia coli* cytosine is only methylated within the sequence 5′-C^{m}CWGG (W = A or T), which is called Dcm-methylation. The status of methylation of this sequence can be analyzed with the isoschizomer *Eco*RII *I-Bst*NI (Table 31.3). While *Eco*RII does not cut the methylated sequence (C^{m}CWGG), the isoschizomer *Bst*NI cuts both the methylated and the unmethylated sequence. Some enzymes are sensitive to Dcm methylation; their consensus sequences are listed in Table 31.3.

In prokaryotes methylation of adenine can be found in the sequence 5′-G^{m}ATC, which is termed Dam-methylation. To analyze this adenine methylation the isoschizomers M*bol*/*Sau*3A can be used (Table 31.3). Both enzymes recognize the sequence GATC. *Mbo*l is sensitive for ^{m}A and the methylated sequence is not cut. In contrast *Sau*3A is insensitive for this methylation and cuts the sequence G^{m}ATC. The methylated G^{m}ATC-sequences can also be detected with *Dpn*l. Interestingly, the enzyme *Dpn*l only cuts the DNA when adenines on both strands of its recognition site are methylated.

It should be considered that because of Dcm- and Dam-methylation the cloning of certain DNA fragments from *Escherichia coli* is sometimes problematical when certain restriction sites are methylated – especially when the methylation motif is encoded through the flanking sequence. The restriction enzymes *Cla*l (ATCGAT) and *Xba*l (TCTAGA) do not cut the DNA when an adenine has been modified by an overlapping Dam methylation (ATCG^{m}ATC respective TCTAG^{m}ATC). The enzyme Stul (AGGCCT) can be inhibited by an overlapping Dcm-methylation (AGGC^{m}CTGG).

Thus it is important to consider flanking bases when a specific recognition site is not cut by the appropriated restriction enzyme. This problem can be avoided by utilizing a methylation insensitive isoschizomer or an *Escherichia coli* strain that is negative for Dcm- or Dam-methylation.

31.4 Methylation Analysis by Methylcytosine-Binding Proteins

This method utilized specific proteins that exhibit high affinity binding to methylated DNA. Different proteins have been isolated from mammalian cells, which can bind methylated cytosine and are involved in the inactivation of gene expression and changes of chromatin state. These proteins are termed methyl-CpG-binding proteins (MeCPs) and possess a methyl binding domain (MBD). Several different proteins (i.e., MeCP2, MBD1, MBD2, and MBD3) have been characterized and isolated. These MBD proteins are used for different methods to analyze the methylation status of specific genomic regions or to identify novel differentially methylated regions (Figure 31.10). For example, tagged MBD proteins are immobilized to nickel agarose beads and used for enrichment of methylated DNA. Genomic DNA is fragmented by sonication or restriction digestion and purified on a column. At a specific salt concentration the methylated DNA is bound to the column, and the unmethylated DNA is eluted. Afterwards the methylated DNA is eluted with high salt concentration and analyzed. Subsequently the methylated DNA can be quantified by real time-PCR, hybridized on microarrays, or analyzed by deep sequencing.

The methylated-CpG island recovery assay (MIRA) uses the ability of MBD3L binding to MBD2 and thereby increases the affinity of MBD2 to methylated DNA. In MIRA recombinant GST-tagged MBD2b protein and His-marked MBD3L1 protein is expressed in bacteria and is purified with glutathione-Sepharose 4B or the respective Ni-NTA agarose beads. The genomic DNA is isolated and cut with the enzyme *Mse*I. *Mse*I recognizes the TTAA cutting site but cuts rarely in CpG islands; 1 µg purified GST-MBD2b and 1 µg His-MBD3L are pre-incubated together with 500 ng

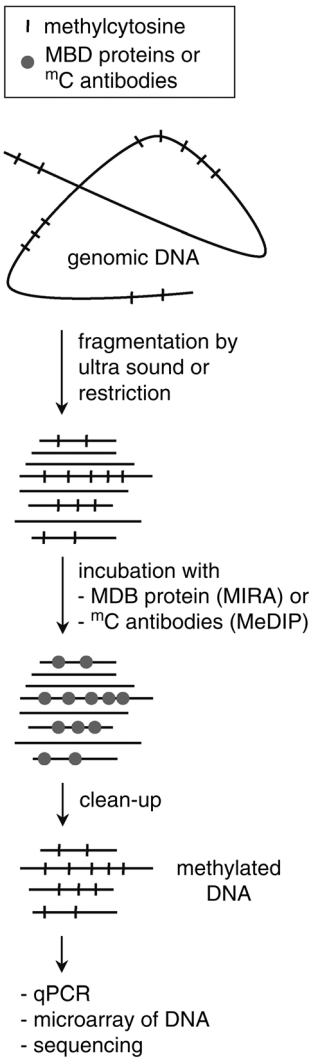


Figure 31.11 Analysis of differential methylation levels of normal and tumor tissue. Genomic DNA is fragmented and purified with MBD proteins (MIRA) or methylcytosine specific antibodies (MeDIP). The methylated DNA is marked by fluorescence and analyzed on microarrays.

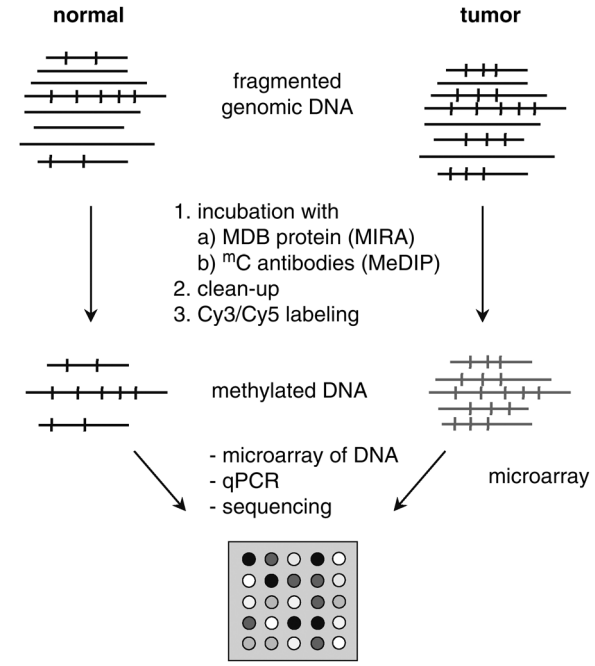


Figure 31.10 Analysis of the DNA-methylation with methyl-binding proteins. The genomic DNA is fragmented by ultrasound or restriction digestion in small fragments and bound specifically to methylcytosine binding proteins (MBD) or to methylcytosine antibodies, precipitated, and purified. The methylated DNA can be quantified by PCR, sequenced, or analyzed on microarrays.

unmethylated DNA (e.g., bacterial DNA) and then incubated for some hours with approximately 500 ng of fragmented genomic DNA. During this step the methylated DNA is bound to MBD2b/MBD3L. The methylated DNA is precipitated with immobilized glutathione paramagnetic particles and the beads are washed. Subsequently, the methylated DNA is eluted from the beads and analyzed. For example, after linker ligation the enriched DNA can be amplified, labeled, and hybridized on microarrays. As an example application, tumor specific DNA methylation can be revealed by fluorescence marking of DNA from tumor tissue with Cy5 (red) and compared to DNA from normal tissue, which was marked with Cy3 (green) (Figure 31.11).

With the chromatin immunoprecipitation (ChIP) method endogenous MBD proteins are linked with formaldehyde to genomic DNA *in vivo*. Subsequently, the protein–DNA complexes are purified and precipitated with anti-MBD-antibody (see also Section 31.7). Again the methylated DNA can be analyzed by PCR, next generation sequencing (NGS), or microarray (ChIP on chip). These methods allow the identification of methylated DNA-sequences in a certain genome.

31.5 Methylation Analysis by Methylcytosine-Specific Antibodies

A further method for detecting and quantifying modified DNA is based on the development of antibodies binding specifically to ^{5m}C, ^{5hm}C, ^{5f}C, or ^{5ca}C (Figure 31.10). These antibodies interact with single strand modified DNA. For example, methylated DNA can be precipitated and analyzed. The sensitivity of the anti-5-methylcytosine antibody is very high. The monoclonal mouse ^{5m}C-antibody interacts specifically with methylated DNA in 10 ng genomic DNA that contains only 3% ^{5m}C. With these antibodies methylated DNA can be precipitated and analyzed by real time-PCR, deep sequencing, or microarray-technology (Figures 31.10 and 31.11). This method is also called the methylated DNA immune-precipitation (MeDIP) method. During the MeDIP procedure 4 µg of genomic DNA is cut into 300–1000 bp fragments by sonication and denatured by heating. An aliquot of the sheared DNA can be used as an input control. Subsequently, the DNA is incubated with 10 µl monoclonal mouse ^{5m}C antibody at 4 °C for several hours. The methylated DNA is precipitated with anti-mouse IgG beads,

purified, and eluted through a proteinase K digestion. The enriched methylated DNA can now be investigated by real-time PCR, sequencing, or microarray analysis. The ^{5m}C, ^{5hm}C, ^{5f}C, or ^{5ca}C antibodies together with immunofluorescence can be utilized to investigate chromosomal segments with a high occurrence of modified cytosines.

31.6 Methylation Analysis by DNA Hydrolysis and Nearest Neighbor-Assays

The following methods allow us to analyze the frequency of modified bases in the DNA and for the nearest neighbor-assays their context as dinucleotides can be detected (e.g., ^{5m}CpN or ^{5hm}CpN). However, with these methods, it is not possible to locate the modified base in the genomic sequence. Since the DNA of contaminated organisms can influence the analysis results, the investigated cells should be free of foreign DNA from viruses, mycoplasms, or other endoparasites. With the DNA hydrolysis method the DNA is completely hydrolyzed. Subsequently, the base composition is fractionized and the modified bases are quantified. Since products of a chemical hydrolysis are rather complex, the enzymatic hydrolysis is the preferred method. Spleen phosphodiesterase or *Micrococcus*-nuclease produces 3′-phosphorylated mono-nucleosides. Pancreatic DNase I or snake venom phosphodiesterase produces 5′-phosphorylated mono-nucleosides. Afterwards, 3′- or 5′-phosphates are eliminated with an alkaline phosphatase and the hydrolysis products are identified by different techniques such as high performance liquid chromatography (HPLC), mass spectrometry, or capillary electrophoresis (CE). With HPLC it is possible to detect in 2.5 µg DNA from 0.04 up to 0.005% ^{5m}C.

With the nearest neighbor analysis the frequency of methylated bases in context of the 3′-neighbor can be dissected. For this purpose the purified genomic DNA is marked with one of four [α-^{32}P]dNTP by nick translation at accidental strand breaks that can be generated by DNase I (Figure 31.12). Then the DNA is digested with *Micrococcus* nuclease and calf thymus phosphodiesterase (exonuclease) to 3′-dNMP, with the radioactive ^{32}P of the 5′-position of the marked nucleotide located now at the 3′-position of the 5′-base. By adsorption chromatography or HPLC the 3′-marked dNMP are separated and then compared to different modified standards. For example, one can analyze how often ^{5m}C or ^{N6m}A is found with respect to one out of four different 5′-bases. Since this method was rather complex, it has been refined. DNA is digested with a restriction enzyme and then the cutting site is marked with a specific [α-^{32}P]-dNTP and Klenow. To analyze methylated CpG, the DNA can be digested with *Mbo*I (/GATC) and the cutting site can be labeled with [α-^{32}P]-dGTP and Klenow (Figure 31.12). With nucleases the DNA is digested to 3′-marked-dNMP and the modified base can be quantified with chromatography. The intensity of labeled ^{5m}dCp, dCp, dTp, dGp, and dAp indicates the quantity of ^{5m}dCpG, dCpG, dTpG, dGpG, and dApG, respectively, at *Mbo*I cutting sites. Alternatively, the DNA can be digested with *Fok*I (GGATGN$_{9–13}$) and marked with one of the four [α-^{32}P]-dNTPs.

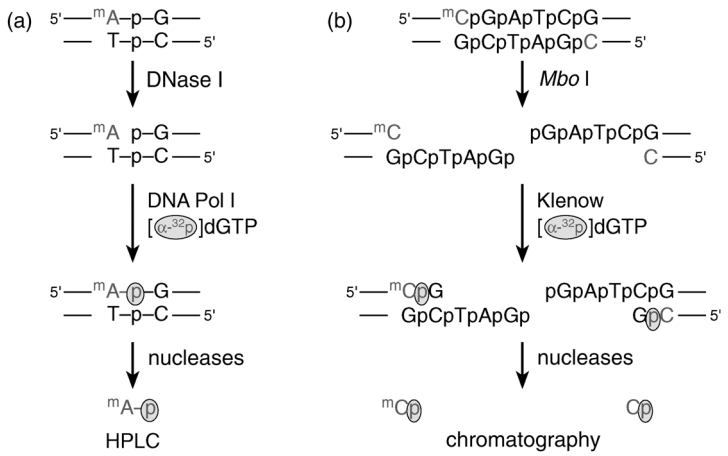


Figure 31.12 Principle of the nearest neighbor-analysis. (a) α-^{32}P-dGTP is inserted at the 3′-end of the methylated adenine and analyzed by chromatography. (b) The DNA is digested with *Mbo*I, the cutting site is labeled with α-^{32}P-dGTP and Klenow. Subsequently methylated Cs are quantified.

The advantage of *Fok*I is that DNA modifications can be investigated independently from their recognition site but in the context all four 3′-downstream bases (NpA, NpG, NpT, and NpG). With the enzyme *Mvo*I (CC/WGG) only the methylation of CpA and CpT can be analyzed. In principle, several enzymes could be used; however, it should be considered that the selected restriction enzyme is not sensitive to DNA modifications. The nearest neighbor-analysis is a preferred technique to identify new DNA modifications; however, with this method these modifications cannot be localized within the genome.

Chromatin Immunoprcipitation (ChIP), Chapter 32.4.7

31.7 Analysis of Epigenetic Modifications of Chromatin

Chromatin modifications and the specific binding of proteins (e.g., transcription factors) to DNA are preferably investigated with chromatin immunoprecipitation (ChIP) (Figure 31.13). This method is based on the crosslinking of nucleic acids with proteins or by crosslinking from proteins among each other with formaldehyde. The crosslinked protein–nucleic acid complexes are precipitated with a specific antibody against the protein or the chromatin modification. The precipitated complexes are washed stringently to elute unspecific bound chromatin. By heating the crosslinks are made reversible and the proteins are digested. The purified DNA can be investigated with PCR (real-time PCR), deep sequencing (ChIP Seq) or microarray (ChIP-on-chip). One important requirement for this method is that the antibody for the analyzed modification or protein works also in the crosslinked chromatin condition.

For ChIP analysis, DNA binding proteins are crosslinked *in vivo* by 1% formaldehyde. During this procedure proteins are linked covalently among each other as well as with the genomic DNA. This crosslinking can be carried out in cell culture for 10 min at 37 °C. Afterwards the cells are washed and harvested. The cells are lysed in a SDS buffer and the chromatin is sheared by sonication into fragments of approximately 500 bp. Subsequently, the chromatin is diluted in a binding buffer to 200 up to 300 µg µl^{-1} protein and an aliquot is retained as an input-control. To reduce the unspecific binding of antibodies the chromatin sample is pre-incubated with protein A agarose. Then the chromatin is incubated at 4 °C overnight with a specific antibody against the chromatin modification or factor (Figure 31.13). As a negative control in parallel an unspecific mouse IgG antibody can be used or as positive control a histone H3 antibody is suitable. The bound chromatin is precipitated by addition of protein A agarose for 1 h at 4 °C and centrifuged. Subsequently, all samples are intensely washed. Finally, crosslinking of samples and the input-control are reversed by the addition of 5 M NaCl and incubation at 65 °C overnight. Proteins are digested by proteinase K and DNA is purified. The enriched DNA can now be analyzed by real-time PCR, ChIP-on-chip (chromatin immunoprecipitation on microarray) or ChIP-Seq (chromatin immune-precipitation with ultra-deep *s*equencing).

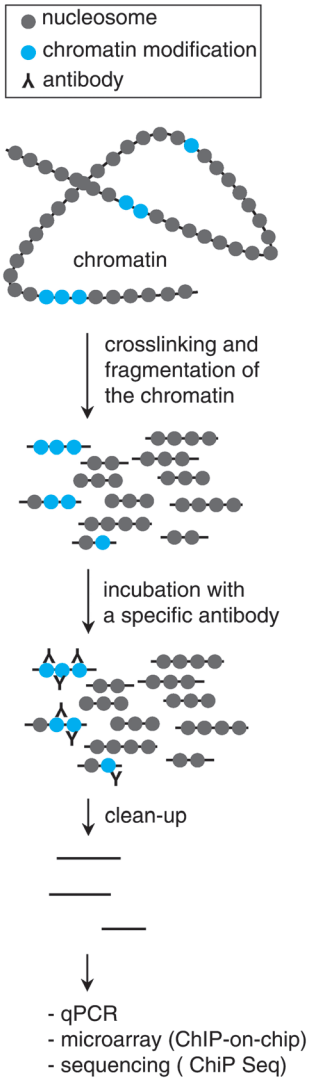


Figure 31.13 Analysis of DNA-associated protein modifications or factors with chromatin immunoprecipitation (ChIP). Chromatin is crosslinked with formaldehyde and fragmented by sonication in approximately 500 bp. The chromatin is incubated with specific antibodies, precipitated, and purified. The purified DNA can be quantified by PCR, sequenced (ChIP-Seq), or analyzed on microarrays (ChIP on Chip).

31.8 Chromosome Interaction Analyses

During the interphase chromosomes are uncoiled and organized in different topologically associating domains (TADs) and chromosomal territories. In this configuration chromatin interacts intrachromosomally as well as interchromosomally. For example, there are intra-chromosomal interactions between enhancers and promoters. This chromosomal topology can be investigated with the chromosome conformation capture technique (CCC or 3C analysis) (Figure 31.14). During the 3C analysis chromatin is crosslinked, cut, and ligated. During this treatment a preferential ligation of associated and crosslinked DNA fragments occurs and these specific interactions can be analyzed by quantitative PCR.

For the 3C method the chromatin is incubated with formaldehyde to covalently crosslink protein–protein and protein–DNA interactions. This treatment can be performed in cell culture with formaldehyde. The cells are lysed and the crosslinked chromatin is cut with a restriction enzyme. Subsequently, the restriction enzyme is deactivated and the chromatin is ligated by T4 DNA ligase. During this step all DNA fragments that are fixed by crosslinking are preferentially

ligated (Figure 31.14). Afterwards crosslinks are reversed by an incubation at 65 °C overnight and the DNA can be purified and precipitated. The amount of ligation products can be analyzed by quantitative real-time PCR. The more often ligation products are detected the higher is the probability that DNA fragments interact *in vivo*. With the help of circular chromosome conformation capture (4C- or circular 3C-analysis) new chromosomal interactions can be detected (Figure 31.14). In this analysis the ligation products resulting from the described 3C-analysis are cut by a further restriction enzyme and transformed in a circular DNA (circular DNA) with a second ligation. These circular ligation products can be amplified by inverse PCR and the unknown DNA sequence can be identified by a sequencing or microarray technique. For this analysis anchor primers for a fixed so called anchor point are utilized.

With the inverse PCR the sequence of unknown circular DNA can be detected. In this respect, a PCR from the known DNA fragment (bait-DNA) into the unknown ligation product is performed.

Hi-C is another modification of 3C that can identify chromatin interactions in a genome-wide manner. Again cells are fixed with formaldehyde and DNA is cut with a restriction enzyme. Subsequently, 5′ overhangs are filled in with biotinylated nucleotides and Klenow. Then a blunt-end ligation is performed under very dilute conditions. This results in a library of ligation products that represent interacting genomic regions. This library is sheared, and the junctions are pulled-down with streptavidin beads. Finally, interacting regions can be identified by paired-end sequencing. Another technique combines ChIP and chromosome configuration capture with high throughput sequencing. This technique, called ChIA-PET for chromatin interaction analysis by paired-end tag sequencing, allows the identification of genome wide chromosome interactions that are associated with a specific chromatin factor or modification.

31.9 Outlook

The DNA modifications and chromatin play important roles in the regulation of gene expression (epigenetics). Since the sequence of the human genome has been elucidated, the DNA sequence is known. The next challenge is to decode the epigenetic modifications and configuration of the chromatin. Tissue- and illness-specific epigenetic patterns could be utilized as biomarkers for early diagnostics of diseases and their molecular classifications.

Further Reading

Beck, S. and Rakyan, V.K. (2008) The methylome: approaches for global DNA methylation profiling. *Trends Genet.*, **24**, 231–237.

Collas, P. (ed.) (2009) *Chromatin Immunoprecipitation Assays*, Methods and Protocols Series: Methods in Molecular Biology, Volume 567, Springer.

Dammann, R., Li, C., Voon, J.H., Chin, E.L., Bates, S., and Pfeifer, G.P. (2000) Epigenetic inactivation of a RAS association domain family protein from the lung tumour suppressor locus 3p21.3. *Nat. Genet.*, **25**, 315–319.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.

Esteller, M. (ed.) (2005) *DNA Methylation: Approaches, Methods and Applications*, CRC Press, Boca Raton.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., Chew, E.G., Huang, P.Y., Welboren, W.J., Han, Y., Ooi, H.S., Ariyaratne, P.N., Vega, V.B., Luo, Y., Tan, P.Y., Choy, P.Y., Wansa, K.D., Zhao, B., Lim, K.S., Leow, S.C., Yow, J.S., Joseph, R., Li, H., Desai, K.V., Thomsen, J.S., Lee, Y.K., Karuturi, R.K., Herve, T., Bourque, G., Stunnenberg, H.G., Ruan, X., Cacheux-Rataboul, V., Sung, W.K., Liu, E.T., Wei, C.L., Cheung, E., and Ruan, Y. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58–64.

Herman, J.G., Graff, J.R., Myohanen, S., Nelkin, B.D., and Baylin, S.B. (1996) Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 9821–9826.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M.,

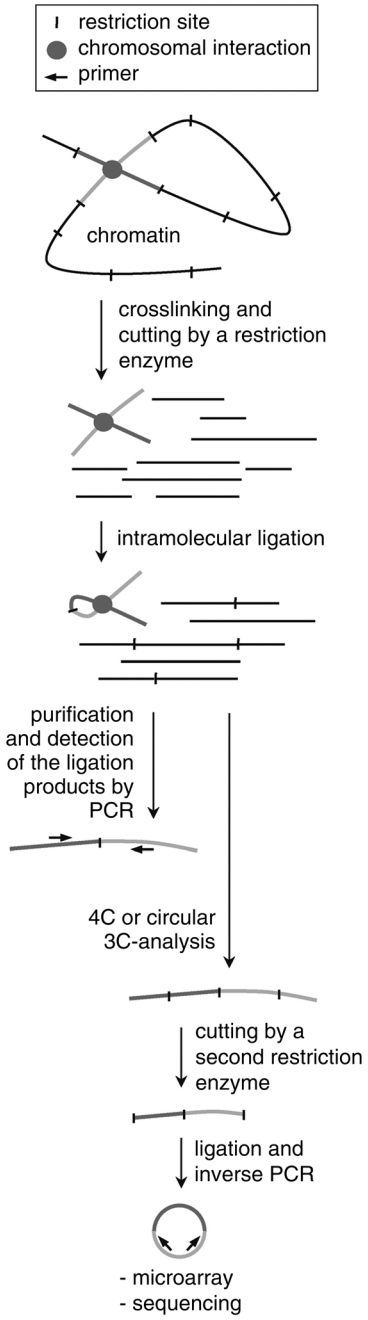


Figure 31.14 Analysis of chromosomal interactions by the *chromosome conformation capture* technique. The chromosomal interactions are fixed by formaldehyde crosslinking. Chromatin is cut by a restriction digestion and the DNA fragments are ligated. The ligation products are purified and analyzed by quantitative PCR (*real-time PCR*) or by other techniques (e.g., inverse PCR).

Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., and Dekker, J. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B., and Ecker, J.R. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

Xiong, Z. and Laird, E.W. (1997) COBRA: a sensitive and quantitative DNA methylation assay. *Nucleic Acids Res.*, **25**, 2532–2534.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., and Ohlsson, R. (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and inter-chromosomal interactions. *Nat. Genet.*, **38**, 1341–1347.

Protein–Nucleic Acid Interactions

32

Rolf Wagner

Gustav-Stresemann-Sraße 15, 41352 Korschenbroich, Germany

Protein–nucleic acids interactions are fundamental for all events in living organisms that serve the conservation and propagation of genetic information.

All steps during the flow of genetic information, such as replication, transcription, translation, as well as events during chromatin remodeling, repair, maturation, or transport are characterized through extensive contacts between nucleic acids and diverse classes of proteins. Notable examples of such proteins are polymerases, transcription factors, helicases, topoisomerases, ligases, or ribosomes and telomerases, the latter representing itself complexes between RNA and proteins.

Hence, for modern molecular biology or biochemistry it is of central importance to unravel the molecular mechanisms underlying the recognition between proteins and nucleic acids.

Although the basic molecular details in the recognition between proteins and DNA or RNA share many common principles there are several subtle differences caused by the fundamental different higher-order structures and functions of both nucleic acid classes leading to specific peculiarities in their interaction mechanisms.

A separate chapter is therefore devoted to describe methods for the analysis of RNA–protein complexes. It should be emphasized, however, that many of the methods described are similarly suitable for the analysis of RNA–protein complexes as well as DNA-protein complexes. This is especially true for the physical methods described in this chapter.

32.1 DNA–Protein Interactions

32.1.1 Basic Features for DNA–Protein Recognition: Double-Helical Structures

DNA predominantly exists as a double-stranded helical structure, which over large ranges consists of the B-form structure as postulated by Watson and Crick. This helical structure is characterized by two base-paired polynucleotide strands, which are intertwined plectonemically by two strands of opposite polarity (strands cannot be pulled apart without unwinding). In this structure, the negative charges of the sugar-phosphate backbone point outwards at an optimal distance. The helix is furthermore characterized by a major groove and a minor groove, which are wound in right-handed turns around the helix axis. The paired aromatic bases (A:T and G:C) are stacked parallel on top of each other perpendicular to the helix axis (tilt). Neighboring base pairs are twisted relative to each other by 36° in right-hand turns (twist), which results in a full helical turn after about ten base pairs. Donor and acceptor positions of the nucleotide bases are involved in base pairing within the helix and thus shielded by the sugar-phosphate backbone from functional side groups of potential outside proteins. Apart from the interactions of polymerases and some single-strand binding proteins the recognition of most DNA binding proteins occurs without breaking the double-stranded base pairing structure. Hence, the interaction between DNA and

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

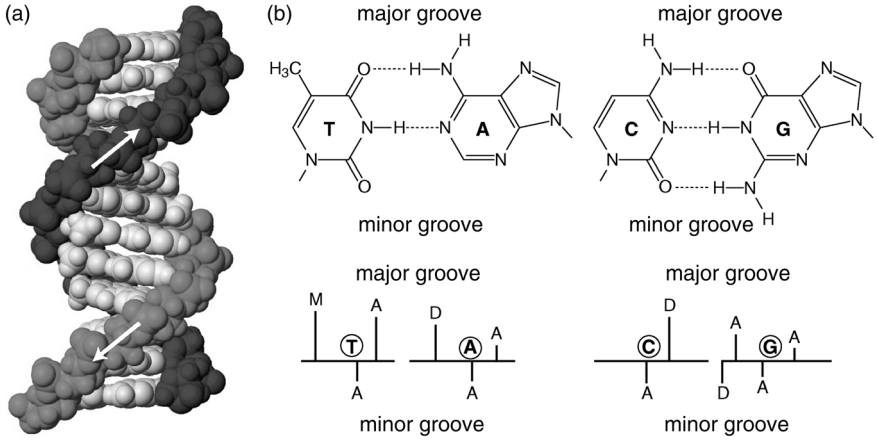


Figure 32.1 Structure of helical B-DNA. (a) Arrangement of the sugar-phosphate backbone and the major and minor grooves. (b) Chemical recognition motifs within the grooves. A: H-bond acceptors, D: H-bond donors, and M: methyl group.

proteins does not involve specific Watson–Crick-type base pairing, which is otherwise extremely important for nucleic acid interactions during biological processes. However, the grooves of the DNA double helix provide very specific surfaces for the recognition of protein structures in which each specific base pair exhibits an individual pattern of H-bond donors, -acceptors, or methyl groups for the interaction with amino acid side chains of proteins. The helical grooves of the DNA play a predominant role in the interaction between DNA and proteins (Figure 32.1).

There are of course special proteins that recognize less abundant DNA structures, such as single-stranded DNA or alternative helix structures, like Z-DNA, for instance, which represents a left-handed structure. Interestingly, these proteins often share structural similarities with RNA-binding proteins.

32.1.2 DNA Curvature

In total, the structure of DNA does not only represent DNA in the B-form. In fact, the exact helical geometry of a given DNA results from the sequence of the different base pairs. Depending on the individual sequence the DNA does not uniformly follow a B-form structure but local differences in the helical conformation occur as a result of deviations in rotational (*twist, tilt*, or *roll* angles) or translational parameters (*shift, rise, slide*) of the base pairs. Often these structural deviations cause a curved path (*curvature*) of the otherwise straight DNA conformation. Such changes in the DNA contour often provide additional recognition signals for the interaction of specific proteins. Characteristically, curvature arises at consecutive A:T base-pairs clustered in helical phase. Yet, DNA curvature can also result from GGCC sequence repetitions. There is, however, a difference in the direction of the curvature induced by A:T clusters or GGCC repeats. At A:T clusters the minor groove points to the inside of the curvature whereas GGCC repeats result in a curvature with the minor groove pointing to the outside of the curvature. The resulting contour angles are quite remarkable and for a single A:T cluster angles between 12° and 23° have been noted. Several models have been put forward to explain the occurrence of curvature. The most descriptive are probably the *ApA-wedge*- and the *B-junction* models (Figure 32.2). In a simplified way, the ApA-wedge model predicts that each A:T dinucleotide causes a change in the tilt and roll angles. This results in a widening of the stacked bases similar to a wedge. Several such alterations in helical phase (ten base-pair distance corresponding to one helical turn) lead to a continuous DNA curvature. In the B-junction model the curvature is explained by the fact that only in the B-form DNA are the stacked bases perpendicular to the helix axis while in other helical conformations, such as the A-form, the plane of the bases relative to the helix axis is changed by the tilt angle. DNA sequences consisting of A:T clusters have a tendency to exist in the DNA A-form. A kink of the helix axis occurs at the junction between the A- and B- conformations because the stacking properties of the aromatic bases force all base pairs to remain packed in a parallel manner. Notably, DNA curvature must not always be static to enable the interaction of curvature-dependent proteins. Often, anisotropic flexibility of the DNA (preferential bending in one direction) suffices to cause a shortened persistence length, facilitating the deformation in one but not the other

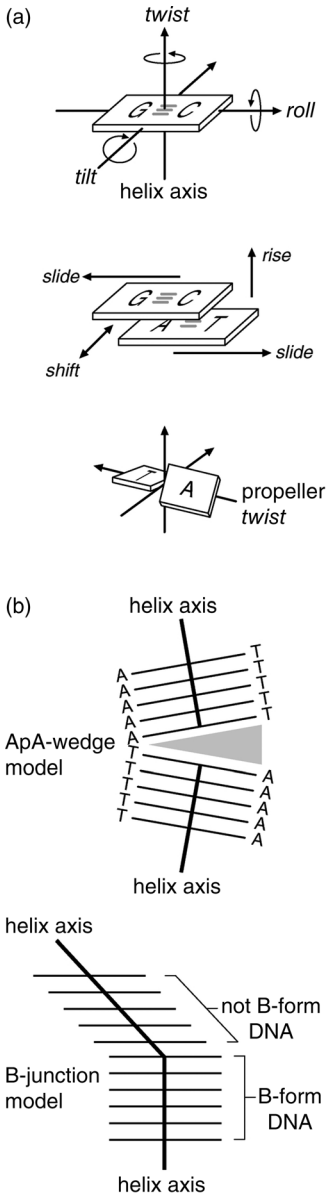


Figure 32.2 Parameters describing helical DNA conformations. (a) DNA-helix parameters; (b) schematic models explaining DNA curvature.

direction. A specific interaction is supported if an adequate adaptation of the DNA structure to the protein surface is possible. This type of induced conformational change is termed DNA bending. DNA curvature, and the concomitant efficiency to bind various proteins, depends on several external parameters including temperature (normally DNA curvature melts above 50 °C) or the presence of bivalent ions, such as Mg^{2+} and Ba^{2+}, which generally enhance curvature, while antibiotics like distamycin, which binds into the minor groove of A:T-rich sequences, cause a reduction in curvature. Moreover, superhelicity has a profound effect on DNA curvature.

How can the curvature of DNA be detected or even the position and intensity of DNA curvature be determined and how could one show that binding of a protein might alter the curvature of a given DNA? The simplest approach applicable in almost every laboratory is gel electrophoresis. As outlined below, DNA curvature reduces the gel electrophoretic mobility. In other words, the speed of migration of a curved DNA fragment is slower compared to a non-curved DNA fragment of the same length. To determine the degree of curvature of a particular DNA one has to compare the gel electrophoretic mobilities of the curved (μ_{obs}) and the non-curved (μ_{act}) DNA fragments. This can easily be done for DNA fragments between 100 and 500 bp with native polyacrylamide gels between 8% or 10%. To measure the difference in mobility the gel electrophoresis has to be performed at low (<25 °C) temperature, which stabilizes curvature, and at a high (>50 °C) temperature at which the curvature but not the base pairs of the double strands melts. The ratio of the two mobilities is termed the *k*-factor ($k = \mu_{act}/\mu_{obs}$). A *k*-factor larger than 1 ($k > 1$) indicates that the DNA is curved, whereby the magnitude of the *k*-factor correlates with the angle of the curvature. Moreover, the magnitude of the *k*-factor depends on the position of the center of curvature within the DNA fragment, which means that reduction of the mobility caused by curvature correlates with the end-to-end distance of the DNA fragment. In fact, the end-to-end distance of a curved DNA is a function of the angle of the curvature and the position of the center of the curvature with respect to the fragment ends. Hence, for the end-to-end distance of a curved DNA fragment the position of the center of curvature matters. For the same angle the distance is smaller if the center of curvature is closer to the middle of the fragment than to its ends. This means that DNA fragments of equal size with the same curvature exhibit the lowest gel electrophoretic mobility if their center of curvature is localized in the middle of the fragment. If one determines the mobility of a DNA fragment with a given curvature in the middle of the fragment (μ_M) and compares this mobility with the same curvature localized at the fragment ends (μ_E) it is possible to derive the curvature angle α from the empirical relationship $\mu_M/\mu_E = \cos(\alpha/2)$. This gel electrophoretic measurement not only allows determination of the degree of curvature it also discloses the center of curvature within a certain DNA fragment. Special plasmids have been constructed to clone curved DNA fragments or curved protein binding sites at different positions within a series of DNA fragments of exactly equal length (circular permutation assay). The different DNA fragments are generated by hydrolysis with a set of restriction enzymes for which cleavage sites have been positioned as direct repeats flanking a central cloning site for the uptake of the desired DNA. The combined restriction digest results in a set of DNA fragments of identical size with the inserted DNA at a different distance relative to the fragment ends. A plot of the gel electrophoretic mobility of the different fragments against the distance of the fragment ends in bp yields the center of curvature as the extrapolated position of minimal mobility. The curvature angle can additionally be derived from $\mu_M/\mu_E = \cos\ \alpha/2$ (Figure 32.3).

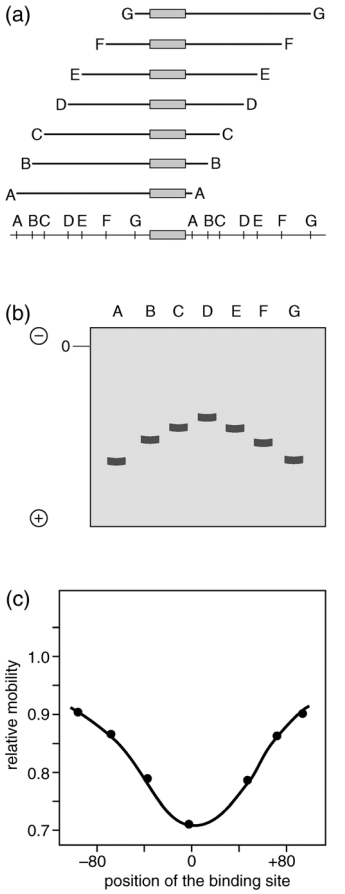


Figure 32.3 Scheme of a permutation analysis to determine the center of curvature within a given DNA. (a) Arrangement of DNA fragments of identical length resulting from restriction hydrolysis with enzymes A–G. Restriction sites of the enzymes A–G for the integration of a DNA-binding region or a curved DNA flank the cloning site (grey box) as direct repeats. Hydrolyses with the different enzymes result in fragments of equal length in which the region for the integration of the DNA in question exhibits different distances to the fragment ends. (b) Schematic depiction of a retardation gel with the different DNA fragments. (c) Diagram showing the relative mobilities of the DNA fragments as a function of the position of the cloning site relative to the fragment start. The center of the curvature is derived by extrapolation of the position with minimal mobility versus the base position.

32.1.3 DNA Topology

Single-stranded DNA forms are extremely rare; hence, most existing DNA conformations can be described by different double-helical forms and sections of static or dynamic curvature. However, biological DNA are often covalently closed circles and/or very large with the ends fixed and not free to move. Such structures give rise to different topological isomers, which are characterized by additional parameters. The topology of DNA molecules is very important for biological processes such as replication or transcription, which generally involve DNA–protein interactions. Some fundamental facts to understand the effect of topology on DNA–protein interaction are listed below.

The spatial description of molecules that exist as closed circles or which are fixed at their ends require an additional dimension defining the topology of the system. Such molecules can exist as different topoisomers. In the case of circular DNA molecules different superhelical structures give rise to different topoisomers. Superhelical structures are divided into *positive* (left-handed screw, DNA overwound) and *negative* (right-handed screw, DNA underwound) superhelical windings. Circular DNA with neither positive nor negative superhelical windings is termed *relaxed*.

For a more detailed description of DNA topology and related phenomena the reader is referred to specialized literature. The parameters relevant for topological molecules are defined by a simple equation:

$$L_K = T_W + W_R \tag{32.1}$$

where L_K is the linking number, T_W is the twisting number, and W_R is the writhing number.

The linking number L_K describes how often a DNA strand is intertwined; L_K represents the topological constant, which can only be changed if a DNA strand is broken; L_K is necessarily an integer. The twist T_W gives the number of rotations of the antiparallel DNA strands around the helix axis. In B-DNA the twist T_W is for instance 10.5 bp per turn. The writhing number W_R reflects the three-dimensional contour of the helix axis and describes the number of superhelical over- or underwindings. For relaxed circular DNA without any superhelical windings $W_R = 0$. For such a molecule the linking number and twisting number are identical ($L_K = T_W$, which follows from $L_K = T_W + W_R$). In the case of right-handed superhelical windings the writhing number is negative ($W_R < 0$), while for left-handed supercoils it is positive ($W_R > 0$) (Figure 32.4).

How can these parameters be influenced? As outlined above, the topological constant L_K can only be changed by breaking covalent bonds (the responsible enzymes in the cell are called topoisomerases). Both W_R and T_W are prone to changes by a number of biological relevant processes, which are related to protein binding. Examples are changes in twist (overwinding or melting of the double-strand structure). Proteins that change the twist upon DNA binding either enhance or reduce the superhelicity of the DNA. In turn, enhancement or reduction of the superhelicity may cause a change in the binding affinity of such proteins. Processes like the intercalation of aromatic amino acids between DNA base pairs change twist automatically. In addition, the intercalation of dyes (e.g., ethidium bromide) or antibiotics, which bind into the grooves of DNA, has a similar effect. The superhelicity also changes if base pars are disrupted and the DNA melts in response to protein binding because unwinding the DNA strands reduces T_W. This effect is valid for all polymerases, which cause DNA-melting over a defined range! Therefore, transcription has a direct influence on DNA superhelicity and vice versa. Note that polymerases, owing to their size and steric constraints, are unable to rotate with the necessary speed (~300 rpm). As a consequence, regions of positive and negative superhelicity flank the section, where RNA polymerase moves (twin supercoiled domain model). Within the cell such regions of superhelical over- or underwound DNA are normally relaxed by cellular enzymes (topoisomerases) (Figure 32.5).

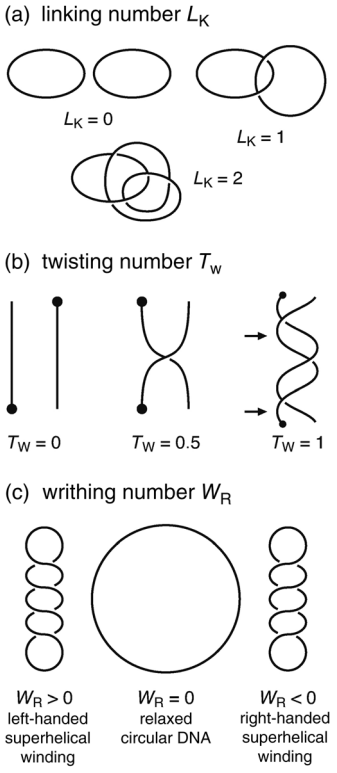


Figure 32.4 Schematic illustration of the parameters L_K (a), T_W (b), and W_R (c) describing superhelical DNA structures.

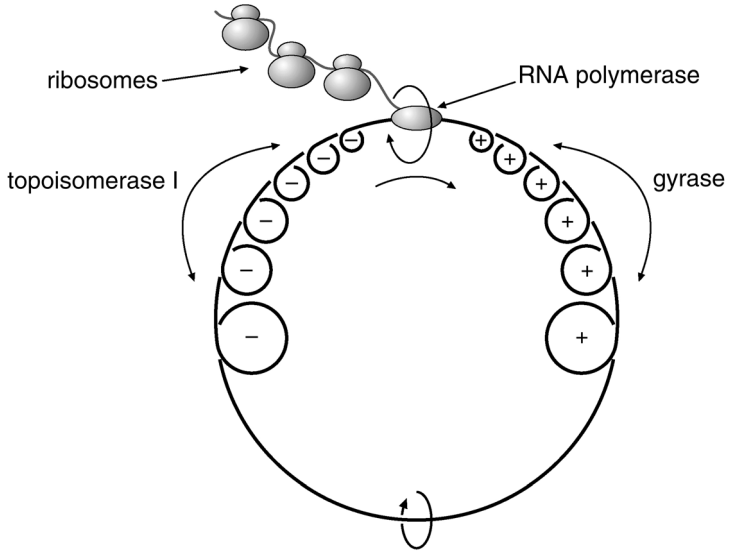


Figure 32.5 Coupling between transcription and superhelical DNA according to the twin supercoiled domain model.

32.2 DNA-Binding Motifs

Comparative structural analyses of known DNA-binding proteins have led to the classification of characteristic amino acid sequence motifs for the recognition and binding of DNA. The most prominent DNA-binding motifs can be divided into five major classes (Figure 32.6): helix-turn-helix structures, leucine zipper structures, zinc-finger domains, helix-loop-helix domains, and β-sheet structures.

Helix-turn-helix structures (HTHs) consist of a section of roughly 20 amino acids in length in which two α-helices are linked by a short β-loop (turn) of approximately four amino acids with an invariant glycine at the second position. Both α-helices are oriented almost perpendicular to each other. The helix closer to the C-terminus is defined as the recognition helix. The recognition helix, which fits exactly into the DNA major groove, is responsible for the recognition. HTH-binding proteins are ubiquitous in prokaryotes and eukaryotes. In prokaryotes HTH-proteins generally recognize palindromic DNA sequences and for that reason normally exist as symmetrical dimers or even-numbered oligomers. Members of eukaryotic HTH-proteins, for instance the homeodomain protein family, bind non-symmetrical DNA sequences as monomers or heterodimers. Some contain additional N-terminal sequences that facilitate binding through the interaction with the DNA minor groove. A variant of HTH-proteins is those containing winged-HTH domains. In winged-HTH-proteins the recognition motif is extended by a third α-helix with a neighboring β-sheet. This secondary structural element makes additional contacts with the DNA backbone.

Zinc-finger proteins exist in many variations and are mainly found in eukaryotes. They all are characterized by a tetrahedral coordination of one or two Zn-ions by conserved cysteines or histidines, which stabilize modular domains of the protein. In the classic case of Zn-finger transcription factors two antiparallel β-sheets, which are linked by a loop with an α-helix, are coordinated by a Zn-ion between two cysteines and two histidines. The DNA contact is maintained by the α-helix, which recognizes a stretch of three base pairs through the major groove. Zinc-finger proteins often consist of multiple such motifs arranged in a consecutive way, such that they are wound helically around the DNA during binding. A special situation is found in Gal4, a yeast transcription factor. Here, two neighboring Zn-ions are bound coordinatively by six cysteines, with each two cysteines sharing one Zn-binding (shared ligands). The two Zn-ions stabilize the position of two α-helices, which also interact with the DNA major groove.

Leucine zipper proteins are designated according to their mechanism of dimerization. They exist as homo- or heterodimers and have almost exclusively been described in eukaryotes. They are composed of an α-helical recognition helix linked to a C-terminal dimerization helix. Dimerization is maintained through hydrophobic interactions between two amphipathic dimerization helices, which form a coiled-coil structure. This structure is characterized by the interaction of each two hydrophobic amino acid residues (generally leucines) separated by two α-helical repeats (heptad repeat) oriented almost on the same site of the helix. The leucine side chains are arranged like the teeth of a zipper. The DNA interaction is maintained by the two separate N-terminal domains that contain positively charged side chains (basic region). These recognition helices are formed like a fork, which fits in opposite directions of the DNA major groove.

Helix-loop-helix proteins (HLHs) are related to zipper proteins. They consist of a shorter DNA binding helix and a longer dimerization α-helix, which is linked by an unstructured loop to a four-helix-bundle. HLH proteins form homo- or heterodimers similarly to leucine zipper proteins. Each one of the α-helixes from the two dimers binds into the DNA major groove. The binding specificity and affinity can thus be modulated by different protein partners.

β-Folds of proteins use their particular secondary structure as the principal element for DNA binding. A pair of anti-parallel β-strands adapts itself into the DNA minor groove. Solved high-resolution structures (e.g., the TATA binding protein (TBP)) reveals that the conserved β-leaf structure of two pseudo-identical domains form a saddle-like structure, which fits into the minor groove of the DNA recognition sequence. Aromatic side chains of two conserved phenylalanines at the end of each β-sheet are intercalated between two DNA base pairs. This interaction creates a DNA kink, such that the DNA points away from the binding protein.

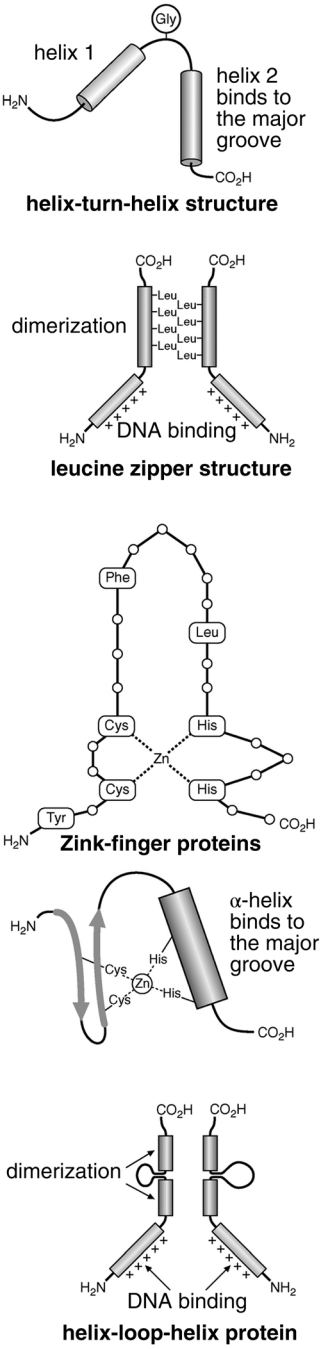


Figure 32.6 Schematic depiction of different DNA binding motifs.

32.3 Special Analytical Methods

Several very powerful methods, starting from very simple to those having a high to very high technical cost, are presented below. As a detailed introduction to the technical and theoretical requirements exceeds the intension of this chapter, different methods will only be introduced briefly and their applications exemplified.

32.3.1 Filter Binding

One of the earliest methods for the analysis of protein–nucleic acid interactions is the filter binding technique, which relies on the principle that proteins bind to nitrocellulose membranes while, for example, nucleic acids, if not too large or complex, migrate through the membrane during a filter process. In a typical binding experiment a mixture of protein and the putative target nucleic acid (preferably radiolabeled) is filtered through a nitrocellulose membrane. The non-bound nucleic acid is subsequently washed from the filter. Existing protein–nucleic acid complexes are retarded on the filter by the protein. The use of radiolabeled nucleic acid allows the amount of complex to be determined by counting the radioactivity on the filter. If adequately performed a differential determination of the filtrate is also possible. Although filter binding is not a real equilibrium binding method it yields relatively exact quantitative data. Filter binding therefore serves to determine apparent binding constants and, moreover, is suitable for slower kinetic measurements. Owing to the low technical requirements and the fact that the method is simple, fast, and generally applicable, filter binding still belongs to the frequently used methods. Note, however, that the mechanism of interaction between a given protein and the nitrocellulose membrane is not completely known. It has been observed that certain proteins are not bound or lose their binding properties by induced conformational changes.

32.3.2 Gel Electrophoresis

EMSA, electrophoretic mobility shift analysis, or short mobility shift or gel retardation describes a method used to separate protein–nucleic acid complexes from the free nucleic acid.

The analysis of protein–nucleic acid complexes by gel electrophoretic methods is, next to filter binding, certainly a relatively easy technical system and probably the most popular method of all for complex analysis. Today, the terms EMSA (electrophoretic mobility shift analysis) or gel retardation summarize qualitative as well as quantitative procedures for the analysis of protein–nucleic acid complexes. The method is equally suitable for the investigation of DNA or RNA complexes. In fact, gel retardation was initially established for the analysis of complexes between ribosomal proteins and ribosomal RNAs while today the technique is preferentially used to study the interaction of DNA-binding proteins. The method is based on the observation that binding of a protein to a nucleic acid generally reduces the electrophoretic mobility of the nucleic acid in non-denaturing polyacrylamide or agarose gels. In a typical experiment the protein or proteins under study are incubated with the nucleic acid and the complexes formed are subsequently separated from the free nucleic acid by gel electrophoresis. Visualization of the complex bands usually occurs by autoradiography of the radiolabeled nucleic acid. If less sensitive methods can be tolerated (above the nanogram range) non-labeled DNA or RNA can also be detected by sensitive staining (fluorescent dyes, ethidium bromide, or toluidine blue). One important advantage of gel retardation for the analysis of protein–nucleic acid interactions is the fact that studies can also be performed with impure protein preparations. Moreover, a binding analysis of several different proteins to the same DNA or RNA molecule is equally possible. Under favorable conditions complexes with different protein stoichiometry can be separated. This is a notable advantage of gel retardation compared to spectroscopic methods. The method requires only minute amounts of material and is applicable in the range of nanograms of proteins or femtomoles (10^{-15} mole) with respect to the nucleic acid. If pure proteins are available thermodynamic or kinetic parameters, such as equilibrium binding constants and association or dissociation rate constants, can be determined (Figure 32.7).

Background to Gel Retardation What are the physical principles valid for the mobility of a DNA molecule during gel electrophoresis? In a first approximation the migration of a DNA

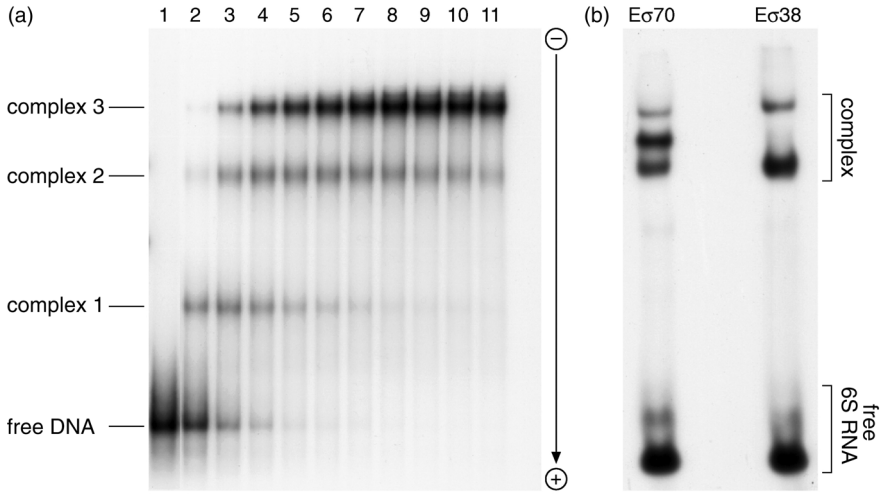


Figure 32.7 Examples of retardation gels. (a) DNA–protein complex analysis. The concentration-dependent binding of the transcription factor FIS to a 260 bp DNA fragment with the regulatory region of the *Escherichia coli rrnD* operon is shown. The FIS concentration is increased in 70 nM steps from 0 to 700 nM (lanes 1–11). The different occupation of three independent binding sites (complexes 1–3) is visible. (b) RNA–protein complex analysis. The binding of a regulatory RNA from *Escherichia coli* (6S RNA) to the bacterial RNA polymerase associated with different sigma factors is shown. The two enzymes form different complexes.

molecule during electrophoresis can be described by the following equation:

$$v = \frac{h^2 QE}{L^2 f} \tag{32.2}$$

where: ν = migration velocity,

h = end-to end-distance of the DNA molecules,

Q = effective charge,

E = electric field,

L = contour length of the DNA,

f = friction coefficient.

DNA molecules, which are generally long and small, behave under the conditions of electrophoresis, where they have to pass through a three-dimensional network of pores, in a worm-like fashion. Their motion can be described by a *reptation* model. The worm-like mobility depends not only on the length of the DNA strand but also on the flexibility (DNA persistence length) and the conformation of the DNA molecules. A static curvature will increase the DNA bulkiness. A worm-like movement of such a molecule through the gel matrix is impeded according to the reptation model. Many proteins induce bends or enhance existing DNA curvature when bound to DNA. Hence, this effect, and not only the change in mass, must be considered in the interpretation of retardation gels.

Electrophoresis, Reptation Model, Chapter 27

For the question of how much the mobility of a DNA–protein complex is altered with respect to the free DNA the mass ratio of DNA and protein is of prime relevance along with the mobility change induced by an altered conformation of the DNA. This ratio, and less the absolute masses of protein or DNA, is especially important for the resolution of complexes during gel electrophoresis. In the case of very acidic (negatively charged) proteins it is possible that no retardation of the complex occurs because the increase in mass is compensated by an overall increase in negative charge by the protein. This has for instance been noted for the Trp repressor protein.

To obtain preferably high resolution between complex bands and free DNA the average pore size of the gel should not be much larger than necessary for separation of the samples in question. The pore size of polyacrylamide gels depends directly on the ratio of concentrations between acrylamide and crosslinker (bisacrylamide) within the gel. At acrylamide concentrations between 10% and 4% the average pore size lies between 5 and 20 nm, respectively, depending on the concentration of the crosslinker. In comparison, the tetrameric Lac repressor has the approximate dimensions of $3.5 \times 3 \times 13\,\text{nm}^3$ and a 50 bp DNA fragment has the approximate dimensions of $2 \times 17\,\text{nm}^2$. For a good resolution, of comparable cases, the acrylamide concentration should not exceed roughly 5%. When choosing the gel system one must also keep in mind that polyacrylamide gels of less than 4% are difficult to handle. For

the separation of larger proteins or longer DNA molecules it is therefore recommended to use agarose gels. The pore size of agarose gels is generally much larger. In the commonly used range of agarose concentrations (0.5–2%) pore sizes of 700–70 nm, respectively, are observed. For very large molecules (polymerases or very long DNA molecules) agarose gels are a good alternative.

One important aspect for the gel analysis of DNA–protein complexes is the stability of the complex during separation. Gel retardation studies benefit from an effect for which the term caging has been coined. This effect is based on the observation that the time for the electrophoretic separation of the samples can exceed the half-life for dissociation of the complex by orders of magnitude. Nevertheless, complex bands can be detected in such cases. The caging effect does not mean that electrophoresis has an effect on the dissociation rate constant. The phenomenon rather results from an enhancement of the local concentration of the reacting molecules, for which the spatial separation, following the dissociation, is reduced in the gel matrix. This enhances the (concentration-dependent) re-association. Moreover, the reduced activity of water molecules within polyacrylamide gels plays an important role in the cage effect. As result of enhanced association kinetics relatively small complex bands can be observed even if the time of electrophoresis exceeds the half-life of complex dissociation several-fold and dissociation may occur several times during separation.

All binding proteins that interact with nucleic acids also undergo non-specific interactions up to a certain degree. Non-specific interactions are essentially the result of different charges of the macromolecules. Generally, non-specific interactions are therefore purely electrostatic. During gel electrophoresis non-specific interactions may give rise to extreme band broadening and impair resolution. Of course, they also interfere with the determination of binding constants.

Specific versus non-specific binding: For all interactions between protein and nucleic acids it is possible to distinguish between non-specific and specific interactions. Non-specific interactions are generally purely electrostatic in nature and result from different charges between proteins and nucleic acids (polyanions). Specific interactions additionally involve H-bonds, hydrophobic interactions, often associated with structural adaptation, stacking interactions between aromatic residues, or directed salt bridges, which contribute to stable binding. Non-specific binding, caused by mere charge interactions, can therefore largely be suppressed by the addition of salt. To relieve non-specific binding during the analysis of protein–nucleic acid complexes it is common, therefore, to add a competitor. Typically, an excess of non-related DNA or the polyanionic compound heparin are added as competitor substance.

It is important, therefore, to distinguish specific from non-specific interactions or better to suppress non-specific binding during analysis. The simplest way is to compensate the surface charge of the interacting partners by the addition of salt (e.g., NaCl or KCl at around 150 mM) to the reaction mixture. The charges contributing to non-specific binding are shielded by the charges of the ions. Yet, the presence of high salt concentrations may affect the efficiency of electrophoretic separations. The general method of choice is the use of an excess of a non-related DNA as competitor substance. As competitor DNA a mixture of chromosomal DNA of different origin, such as calf thymus DNA, may be used. Natural DNA may, as a disadvantage, contain unwanted specific binding sites. The use of a synthetic DNA, like poly(dI-dC) or poly (dA-dT), avoids such potential complications. Heparin, a polyanionic compound (muco-polysaccharide with sulfate groups), has been proven to be of general value as competitor substance. There are no fixed rules predicting the exact amount of competitor substance required for a particular gel retardation experiment. The appropriate concentration has to be determined in prior titration experiments for each different binding partner. A good start is between 20 and 150 ng μl^{-1} competitor in the reaction mixture. Another rule of thumb suggests the use of a roughly 200-fold molar excess of competitor with respect to the target DNA. Notably, however, at high competitor concentrations not only the unbound protein will be captured by reducing the non-specific interactions. It is known that high competitor or heparin concentrations may also actively initiate the dissociation of complexes. For several complex-binding interactions, defined by strongly different on- and off-rates, the order of addition of the competitor must be considered. The choice of a suitable competitor, the time of addition, and the optimal adjustment of its concentration are therefore crucial steps for any gel retardation experiment!

32.3.3 Determination of Dissociation Constants

In a reaction where a protein–DNA complex is in equilibrium between binding and dissociation the equilibrium is defined by the equilibrium constant K. The value of K is given by the ratio between the concentration of the complex and the product of the concentrations of unbound protein and free DNA in equilibrium. The inverse measure $1/K$ is termed the dissociation constant K_D.

To determine dissociation constants one selects a low DNA concentration (smaller than the dissociation constant expected) and adds increasing concentrations of protein to the reaction. If the DNA concentration is much smaller than the dissociation constant, then, under equilibrium conditions between free DNA and DNA–protein complex, the concentration of protein equals the dissociation constant:

$$K_D = \frac{[P][D]}{[PD]} \tag{32.3}$$

where P = protein and D = DNA-binding site. If $[D] \ll K_D$ it follows that $[P]_{free} \approx [P]_{total}$, leading to:

$$K_D = \frac{[P]_{total}[D]}{[PD]} \tag{32.4}$$

For a quantitative determination of complex dissociation constants one normally measures the concentration of protein that is required to bind half of the DNA present. This is best done in a pilot experiment, where the protein concentration is varied over several orders of magnitude. The range of the dissociation constant can be taken from the protein concentration that yields half-saturation of the complex. To obtain a more accurate value of K_D from the visually estimated half-saturation the precise amounts of free DNA and DNA–protein complex are determined by a preferably exact method of quantification. Typically, the amount of free DNA is plotted against the logarithm of the protein concentration (Bjerrum plot). Provided that complexes can be resolved sufficiently and quantitatively determined by exact densitometric methods (autoradiography or phosphoimaging) binding constants can be obtained with relatively high precision. For a correct measurement of the binding or dissociation constant it is important that the ratio of DNA concentration under conditions of half-saturation is small with respect to the protein concentration (0.01–0.1).

The quantitative evaluation can either be determined from the intensity increase of the complex or from the decrease in concentration of the free DNA. Both methods yield consistent results if the separation technique (gel electrophoresis) itself has no effect on the formation. In case of unstable complexes or badly resolved complex bands, which impede an exact quantification, the measurement of the free DNA concentration is often the only possible way to determine the K_D value.

Importantly, dissociation constants measured by gel retardation do not reflect real equilibrium determinations and, therefore, only apparent constants are obtained. These generally depend on the gel electrophoretic separation conditions (e.g., temperature, competitor concentration, buffer conditions). To gain exact binding or dissociation constants the K_D values obtained should be verified by real equilibrium measurements.

Gel retardation also provides a method for determination of the kinetic parameters of a binding reaction. To measure association binding constants DNA and protein samples are mixed and aliquots of the mixture are separated at short time intervals by gel electrophoresis (samples may be loaded on a running gel!). The association reaction is stopped by virtue of the instant separation of free DNA. The amounts of free DNA or formed complexes are determined by densitometry as outlined above. In this way, the increase in complex concentration over time (rate of complex formation) is obtained. If dissociation rate constants are required the experiment starts with pre-formed complexes, which exist in equilibrium. Dissociation is induced by rapid dilution of the sample or addition of excess competitor DNA (quenching). To ensure that dissociated proteins are effectively captured the competitor concentration must clearly be larger than the protein concentration and is generally used in excess. Aliquots of the sample are separated by gel electrophoresis immediately after the quenching reaction. Of course, measurements of kinetic constants by gel retardation are limited by the time scale of sample manipulation and are therefore hardly feasible in the range below a few seconds.

The stoichiometry within a formed DNA–protein complex can be determined in several ways. Double labeling of DNA and protein can generally be applied. In the case of radiolabeling the protein is labeled with [^{3}H] and the DNA usually with [^{32}P]. The maximal energies of the radioactive decay spectra of both isotopes are sufficiently different to allow parallel scintillation measurements. An exact stoichiometry of the complex partners can be derived if the specific activities of both DNA and protein are known. Alternatively, with somewhat lower sensitivity, a quantification of non-labeled proteins can be performed by quantitative Western blotting of a retardation gel with known amounts of DNA or protein staining of the complex band with Coomassie blue.

The gel retardation method also enables the identification of unknown DNA-binding proteins, which interact with a specific target DNA (in cases in which cell extracts or mixtures of proteins are used). Such analyses require to scale-up of the amount of DNA for complex formation compared to the radiolabeled gel retardation assays because in gel retardation the amount of DNA generally limits the amount of complex formed. For a DNA fragment of roughly 200 bp and a given dissociation constant in the micromolar range about 0.5 µg of DNA are required. The amount of protein in question depends on the affinity to bind DNA and in the case of a K_D of 10^{-6} M should be employed in the range of micrograms. The non-labeled DNA and the DNA–protein complex are briefly stained with ethidium bromide. The stained complex band can be excised and the included protein is separated in a second electrophoretic step by a discontinuous SDS gel. If the protein is unknown the stained protein band can be excised from the SDS gel, digested with suitable proteases, and identified by mass spectrometry (MALDI-TOF analysis).

With known proteins, for which antisera might be available, the retarded complex band can be blotted on a membrane and the protein identified subsequently by Western analysis. Often, however, DNA–protein complexes are difficult to transfer to membranes and in unfavorable cases excess free protein migrates at the position of the complex band. In such cases it is sometimes possible to obtain a supershift on the retardation gel by the addition of an antibody, which specifically interacts with the binding protein. A supershift occurs because the complex is additionally retarded by the molecular weight of the antibody. In rare cases it is possible, however, that the antibody obscures complex formation or causes dissociation. Indirectly, this can be taken as an indication for the identity of a protein.

32.3.4 Analysis of DNA–Protein Complex Dynamics

TGGE: temperature-gradient gel electrophoresis; a gel electrophoretic method used to study changes in the gel electrophoretic mobility of nucleic acids or protein–nucleic acid complexes resulting from temperature-dependent conformational changes of the sample.

The thermal stability of protein–nucleic acid complexes can ideally be studied by the method of temperature-gradient gel electrophoresis (TGGE). In this method a horizontal gel chamber is used in which a linear temperature gradient perpendicular to the direction of separation is installed. The method has great potential to study temperature-induced structural transitions of molecules that change the electrophoretic mobility of the sample to be separated. This includes of course, along with the dissociation of complexes, conformational changes of the nucleic acid (or protein) itself. Self-evidently, the method has a large application area for RNA molecules and RNA–protein complexes.

The temperature range at which DNA–protein complexes dissociate is generally below the DNA melting point of roughly 70 °C. This does not apply, however, to DNA curvature or protein-induced DNA-bending. In such cases structural transitions generally occur in a range below 50 °C, which is considerably below the melting point of DNA. If the binding of a particular protein depends on DNA curvature this can elegantly be studied within this temperature range by TGGE analysis (Figure 32.8).

If the DNA curvature is caused by A:T clusters arranged in helical phase the curvature-dependent DNA-binding can alternatively be studied using substances that influence DNA conformation. Among such substances are for instance the antibiotic distamycin. This oligopeptide specifically binds to the DNA minor groove bent by A:T clusters and abrogates existing curvature. The distamycin-dependent dissociation of DNA–protein complexes indicates that the binding mechanism involves DNA curvature. The effect of curvature-removal can be demonstrated directly with the free DNA by TGGE but also by the use of distamycin because the loss of curvature enhances the electrophoretic mobility, which in most cases can be easily recognized by gel electrophoresis. Such studies are especially suitable to analyze the effect of proteins on DNA conformation.

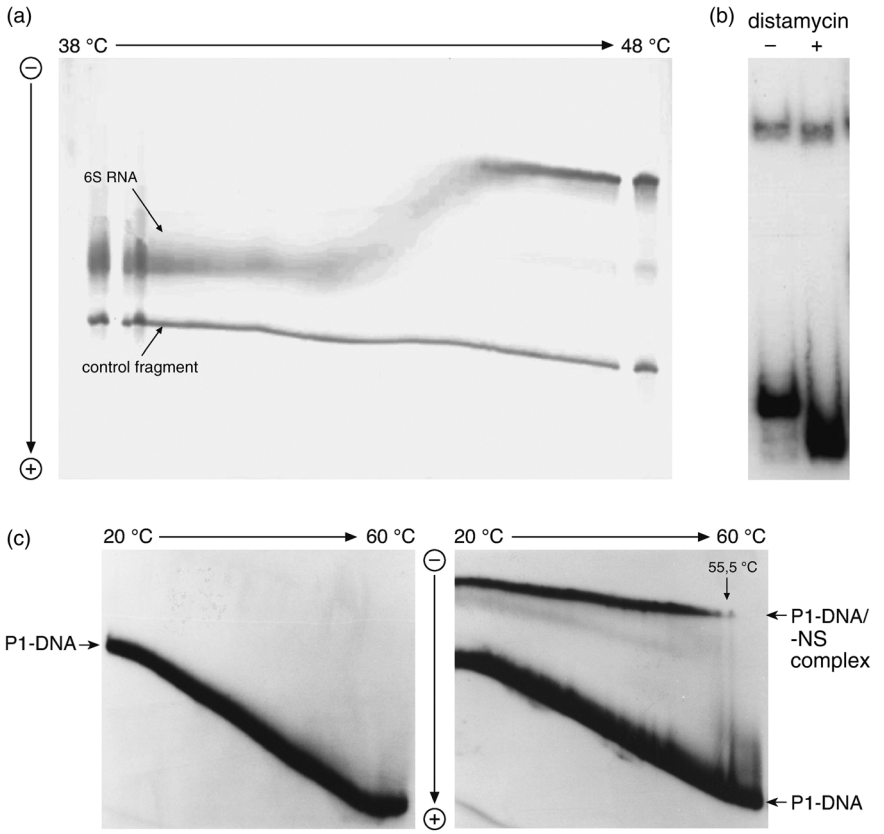


Figure 32.8 Examples of conformational analyses of nucleic acids by gel electrophoresis. (a) Temperature-dependent conformational change of RNA (*Escherichia coli* 6S RNA) analyzed by temperature-gradient gel electrophoresis (TGGE); a 7.5% polyacrylamide gel is shown, a temperature gradient between 38 and 48 °C perpendicular to the direction of electrophoresis has been applied. A clear conformational transition of 6S RNA is evident at 43 °C. A 400 bp DNA fragment was loaded as control. (b) Conformational change of a curved DNA on a native polyacrylamide gel. Shown is a 260 bp DNA fragment (containing the regulatory region of a bacterial rRNA promotor) with intrinsic curvature. The DNA curvature caused by A:T clusters of the sequence is abrogated in the presence of distamycin, which results in an increased gel mobility. (c) TGGE analysis of the thermal stability of a DNA–protein complex. The left-hand panel shows the separation of a 260 bp DNA fragment with the regulatory region of a bacterial rRNA promoter (P1 DNA) with a temperature gradient of 20 to 60 °C. The right-hand panel shows a separation of the same DNA bound to the DNA-binding protein H-NS under the same conditions. It can be seen that the complex disintegrates at 55.5 °C.

32.4 DNA Footprint Analysis

In molecular biology a footprint denotes different accessibilities of protein–nucleic acid complexes towards chemical reagents or limited enzymatic hydrolysis, usually visualized by gel electrophoresis. It serves to localize or to narrow down binding sites between the protein and the nucleic acid.

To characterize the interacting domains of a protein with its target DNA the term footprint analysis has been coined, which indicates that something like an "imprint" of the protein on the surface of the DNA is determined. Footprint analyses are generally performed in combination with gel electrophoresis. They are based on the determination of the accessibility of the nucleic acid towards nucleases or modifying reagents, which differs in the absence or presence of a bound protein. Conversely, the accessibility of a protein towards proteases or modifying reagents can also be determined. Positions at which proteins and the nucleic acid are in contact are naturally less accessible, resulting in a reduction of the enzymatic hydrolysis or chemical modification at this region. Enhanced signals, which may occur as well, are generally interpreted as a conformational change of the binding partners. DNA fragments derived from enzymatic hydrolysis or cleavage at the modified positions are separated at the nucleotide level by denaturing gel electrophoresis. A separate DNA sample hydrolyzed or modified under the same conditions but in the absence of the protein is separated in parallel as control.

The difference in band intensities between both samples indicates the positions where the binding protein has been localized. Yet, footprint analyses can provide much more details of a DNA–protein complex than the information about the sequence region where the protein was bound. Special footprint techniques may yield information about the mode of interaction in addition to indicating the borders of protein contact. Conclusions about which bases or chemical groups of the nucleic acid are involved in binding or whether the protein binds into the minor or major DNA groove are possible, even whether binding induces conformational changes or if the DNA strands are interrupted into single-strands during binding can be inferred from footprint data (Figure 32.9).

Apart from such structural information footprint analyses are suitable to determine dynamic processes, such as conformational changes or they can be used for measurements of binding constants. One advantage of the binding constant determination by footprints versus gel retardation is the fact that footprint analyses can be performed under real equilibrium conditions

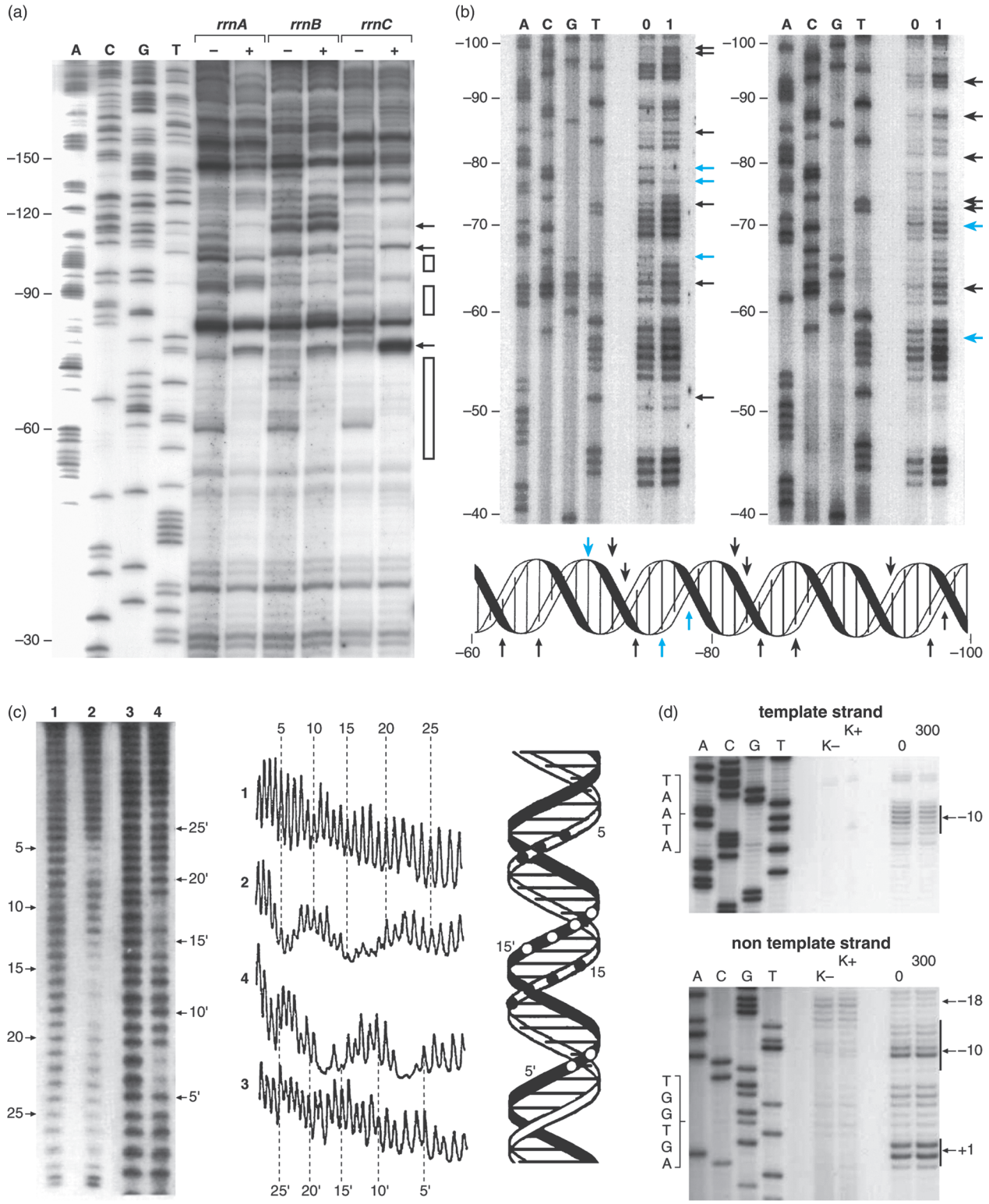


Figure 32.9 (a) Example of a DNase I footprint. The figure shows a section of a primer extension analysis following limited DNase I hydrolysis of the upstream regulatory regions of three rRNA promoters from *Escherichia coli* (*rrnA*, *rrnB*, and *rrnC*) in the presence (+) or absence (–) of the binding protein FIS. Arrows indicate enhanced reactivity (conformational change), open bars mark protected regions. Sequencing lanes are indicated by A, C, G, and T and the sequence positions relative to the transcription start site are shown. (b) Example of a DMS footprint. Results from a primer extension analysis of a DMS footprint of the *rrnB* operon regulatory region are shown. Nucleotide positions and sequencing lanes are indicated. The footprint reaction was performed in the presence (1) and absence (0) of FIS (left-hand side) or presence (1) and absence (0) of H-NS (right-hand side), respectively. Blue arrows indicate protections, black arrows enhanced accessibilities. The lower part of the figure summarizes the reactivity differences in a DNA helix

because protein and nucleic acids are not separated during the reaction. Of course, to solve these specific problems special footprint techniques are required, which will be described below.

32.4.1 DNA Labeling

The simplest and most direct method visualizing footprint bands involves the radioactive labeling of DNA before complex formation and hydrolysis. The resulting fragments of different length are subsequently separated by denaturing gel electrophoresis and their position on the gel is visualized by autoradiography or the use of a phosphoimager. Radioactive labeling of the DNA is only performed at one end of the molecule. This is important because labeling multiple positions results in more than one labeled product band after hydrolysis, which obscures identification. Two principal methods for labeling DNA ends can be used. End-labeling can be performed at either the 5′ or the 3′ ends of each strand. Polynucleotide kinase and γ-[^{32}P]-ATP as substrate are used for 5′ labeling. If the 5′ ends are already phosphorylated, as it is the case after restriction enzyme hydrolysis, the 5′ phosphate group must first be removed by a phosphatase. Alternatively, a kinase reaction may be performed under "exchange" conditions. If DNA fragments generated by PCR are used the 5′ ends are normally not phosphorylated, except phosphorylated primers were employed. Radioactive labeling at the 3′ ends can best be done by a "fill-in" reaction with DNA polymerase provided that the DNA has a 5′ overhanging end. This can be achieved by use of appropriate restriction enzymes. Suitable enzymes for the "fill-in" reaction are either T4 DNA polymerase or the Klenow fragment of DNA polymerase I and the suitable α-[^{32}P]-dNTP. Both the 5′ kinase reaction and also the "fill-in" reaction at the 3′ ends of a DNA molecule can take place at each of the two antiparallel DNA strands. To avoid such double labeling, which obscures the assignment of hydrolytic fragments, a strategy has to be followed that either limits the labeling reaction to one strand or the second labeling has to be removed by use of a single cutting restriction enzyme close to the distant DNA end.

Radioactive Labeling, Labeling Positions, Section 28.3.1

Enzymatic Labeling, Section 28.3.2

32.4.2 Primer Extension Reaction for DNA Analysis

Bands resulting from footprint reactions can be visualized indirectly without the prior use of radiolabeled DNA. A variant primer extension reaction is used for this indirect way of analysis. The principle of the primer extension reaction for DNA corresponds to the reaction that is frequently used for the analysis of RNA molecules.

Primer extension describes a method in which a nucleotide sequence (RNA or DNA) is transcribed into a complementary DNA sequence (cDNA), starting from a DNA oligonucleotide (primer). The primer oligonucleotide is chosen such that it specifically binds to a complementary target sequence located 3′ from the sequence to be transcribed. The primer is subsequently extended in the presence of deoxyribonucleotide triphosphates by Klenow DNA polymerase (for DNA primer extension) or by reverse transcriptase (for primer extension of RNA). The extension reaction either stops at the 5′ end of the target nucleic acid or at bases, which are modified by footprint reactions, which cause the elongation reaction to become blocked (chemical footprint).

During the primer extension of RNA a cDNA is created by reverse transcription of an RNA template. In contrast, for the primer extension method of DNA, a primer oligonucleotide selecting a DNA strand for the copying reaction is used and instead of reverse

model. (c) Example of a hydroxyl radical footprint analysis. The analysis of the binding of Epstein–Barr nuclear antigen (EBNA) to a synthetic consensus DNA sequence is shown. On the left-hand side the autoradiogram of hydroxyl radical cleavage products of the upper (lanes 1, 2) and the lower strand (lanes 3, 4) are shown. Lanes 2 and 4 reflect the presence of the binding protein. In the middle is a densitometric evaluation of the autoradiogram. Lanes 1 and 2 indicate the upper, lanes 3 and 4 the lower strand. Lanes 2 and 4 reflect the presence of the binding protein. The scheme on the right-hand side depicts helical B-form DNA with the upper strand (light) and the lower strand (dark). Protected regions are marked by dots. (d) Example of a $KMnO_4$ footprint analysis for identifying single-stranded DNA regions. The figure shows the primer extension analysis of the $KMnO_4$-reacted template strand DNA of a bacterial promoter (upper panel) and the same analysis of the non-template strand (lower panel). K^- and K^+ mark control lanes without binding protein before and after $KMnO_4$ reaction, respectively. The lanes 0 and 300 indicate the absence or presence, respectively, of RNA polymerase during the reaction, which renders the promoter region into a single-stranded "transcription bubble" giving access to the nucleotides for modification. Source: Part (c), Kimball, A.S., Milman, G., and Tullius, T D. (1989) *Mol. Cell Biol.*, **9**, 2738–2742. With permission, Copyright © 1989, American Society for Microbiology. Part (d), Jöres, L. and Wagner, R. (2003) *J. Biol. Chem.*, **278**, 16834–16843. With permission, Copyright © 2003, The American Society for Biochemistry and Molecular Biology, Inc.

transcriptase a DNA polymerase serves as enzyme. In that case, the DNA fragments obtained by limited hydrolysis or chemical modification of the footprint reaction are directly used as templates for the synthesis of a complementary DNA strand by the Klenow fragment of DNA polymerase I. The polymerization reaction is started by a primer oligonucleotide, which is complementary to a sequence at or close to the 3′ region to be analyzed. Information can be obtained for both strands if appropriate primers are selected, which are complementary to either the upper or lower strand. Generally, primers are always extended in the 3′ to 5′ direction. The synthesis stops at the 5′ ends of the DNA fragments and complementary DNA strands to the fragments resulting from the footprint reaction are generated. Many of the chemical reagents for DNA footprint analysis alter the nucleic acid bases in such a way that the primer extension reaction becomes aborted. Hence, these fragments can also be used directly for primer extension analysis without the need of prior hydrolysis of the DNA strand. To visualize the newly formed DNA fragments either the primer used can be radiolabeled at the 5′ end by a kinase reaction or the extension reaction is carried out in presence of a suitable α-[^{32}P]-dNTP. Because the primer extension reaction works with non-labeled DNA and the sequence region as well as the particular strand can be selected by the primer, footprint reactions of DNA molecules can be analyzed that are normally much too large to be separated on sequencing gels. The gel will only show the labeled products of the primer extension reaction, which can be chosen in a way that is optimal for the resolution of sequencing gels. As a further advantage of the primer extension analysis, circular and superhelical DNA molecules can also be studied, which enables the analysis of DNA–protein interactions of different DNA topoisomers at conditions close to the *in vivo* situation.

32.4.3 Hydrolysis Methods

To obtain information by limited enzymatic digestion, preferably about an entire DNA molecule, sequence-independent DNases are employed. DNases are classified as endonucleases, which can cut anywhere within the molecule or exonucleases, which only digest DNA from either the 5′ or the 3′ end. Sequence-specific restriction enzymes (endonucleases), which are frequently used in molecular biology, are not of particular use for the analysis of protein–DNA complexes by limited hydrolysis because they cut at only a few specific sites.

A basic prerequisite to obtain information by enzymatic hydrolysis over the entire length of a particular DNA molecule is given by the fact that the hydrolysis must be limiting. Principally, each nucleotide position accessible within the sample should statistically be cleaved but each DNA molecule should only be cleaved (hit) once. Such a situation is characterized as a single hit condition. For a DNA molecule of limited length, as generally used in footprint analyses, single hit conditions are fairly easily achieved. This is because the probability of hitting a molecule that has not yet been hit before is much greater than that of hitting the same molecule twice. It can be shown for DNA molecules between 50 and 200 bp that each molecule is statistically cleaved only once if about 70% of the starting molecules remain non-cleaved. Hence, if the limited hydrolysis reaction is performed under conditions that leave about 70% of the starting material non-cleaved, one can assume single hit conditions with statistical distribution of the single cuts. Whether such conditions are met in a particular experiment can easily be inferred when about 70% of the input DNA migrates as non-cleaved material on the gel at the same position as the non-treated DNA sample.

DNase I The nuclease used most frequently for mapping of protein-binding sites within a DNA molecule is DNase I. This endonuclease cleaves DNA largely sequence-independent (there is a slight preference for purine-pyrimidine sequences). On the other hand, it has some structural limitations in the hydrolysis of double-stranded B-form DNA. The enzyme fits specifically into the minor groove of B-form DNA and cleaves the phosphodiester bond such that the phosphate remains at the 5′ end of one of the cleaved strands. The other end is left with a free 3′-OH group. Due to the dimensions of the enzyme a width of the minor groove of about 13 Å is optimal for the hydrolysis reaction. Any distance smaller or larger than this has a negative effect on the enzyme. Owing to this steric requirement the hydrolysis efficiency becomes reduced when the DNA minor groove is reduced by the DNA conformation. This is the case for curved DNA, for instance. The minor groove at the inside radius of the curvature is

only poorly cleaved by DNase I. At sequence regions with regular A:T clusters as well as at some G:C-rich sequences causing an intrinsic curvature the minor groove at the inside of the curvature is so narrow that cleavage by DNase I, due to its dimension, is hindered for steric reasons. This is why, at limited DNase I hydrolysis, sections of strong cleavage alternate with regions of weak hydrolysis signals. DNase I does not therefore create a completely regular hydrolysis pattern of natural DNA along the entire length. DNase I is very efficient, which means that only very small amounts of the enzyme and correspondingly short reaction times are required for a limited hydrolysis reaction. The temperature may also be kept low to avoid too strong a hydrolysis. The enzyme requires Mg^{2+} ions for its activity. This fact can be used to control the hydrolysis reaction, which can precisely be started by the addition of Mg^{2+} ions after all necessary compounds have been added to the reaction mixture (this method of activating DNase I is only possible if the complex formation between DNA and protein does not itself depend on the presence of Mg^{2+} ions). A rapid addition of excess EDTA, which sequesters all the required Mg^{2+} ions, causes the immediate end of the hydrolysis reaction. The enzyme can then be removed by phenol extraction. DNase I can also perform hydrolysis reactions within the gel matrix. This fact can be used to perform footprint reactions of complexes present on retardation gels, which have been used for their separation (in-gel footprint). Complexes separated from the unbound DNA must not be extracted from the gel, which often may result in dissociation. For the in-gel hydrolysis, the complex band is favorably cut from the retardation gel and a DNase I solution in the presence of EDTA is allowed to soak into the gel piece (this is a slow process and for a gel piece 1–2 mm in size requires about 15–30 min). The reaction is then started and stopped by the addition of Mg^{2+} ions or EDTA-containing buffer, respectively (ions diffuse rather rapidly into the gel). The hydrolyzed DNA fragments can subsequently be removed from the gel by extraction in the presence of phenol. After a precipitation step the sample may be analyzed directly on a sequencing gel.

Exonuclease III Along with DNase I several exonucleases are suitable enzymes for footprint reactions. Exonuclease III (Exo III), as indicated by its name, hydrolyzes DNA only from its 3′ end. The hydrolysis starts at the 3′ end and progresses from 3′ in the 5′ direction; hence both antiparallel strands of a DNA fragment are digested in opposite direction (both in the 5′ direction). The preferred substrate for Exo III is double-stranded DNA, which under ideal conditions of complete hydrolysis yields two single-stranded fragments consisting of about half the length of the original DNA, each containing the original 5′ end. Exo III hydrolysis stalls at sites where a protein is stably bound to DNA. Characterization of the remaining DNA fragments after hydrolysis allows definition of the DNA borders protected by the interacting protein. For the analysis only one DNA strand is labeled at the 5′ end. In contrast to DNase I footprint reactions Exo III reactions are not performed under limiting conditions because, rather, as many DNA molecules as possible should be digested until they reach the protein border. In a parallel experiment, in which the opposite strand is labeled, the protein border on the other strand can be mapped, completing the protein binding site information. Some problems with Exo III footprints may be encountered if the protein binds extremely weakly and becomes displaced by the nuclease. Certain stable DNA secondary structures may also disturb the analysis and in cases where a protein binds to multiple sites on the DNA only the one closest to the 3′ end is detected.

λ-Exonuclease Another exonuclease, with complementary direction of hydrolysis, is λ exonuclease. This enzyme degrades DNA from its 5′ end, which makes it a complementary tool to Exo III. For footprint analyses the DNA must of course be labeled at the 3′ ends!

Exonuclease Activity of DNA Polymerases To map the borders of DNA-binding proteins the 3′–5′ exonuclease activities of *proofreading*-competent T4 or T7 phage DNA polymerases may generally be employed. In reactions under footprint conditions any substrate NTP is omitted to favor the exonuclease over the polymerization activity.

In rare cases, where single-stranded DNA should be analyzed, the highly single strand-specific nuclease S1 can be used. For S1 mapping reactions, like for DNase I reactions, strict single hit conditions are required. Nuclease S1 cuts DNA in a sequence-independent manner and, because the specificity of the enzyme is not limited to DNA, S1 is also of value for structural analysis of RNA molecules.

32.4.4 Chemical Reagents for the Modification of DNA–Protein Complexes

Along with the limited enzymatic hydrolysis for the characterization of DNA–protein complexes the use of nucleic acid-modifying chemical reagents is an approved method. It has the advantage that due to the small dimensions of most reagents they can penetrate closer to the interacting surface of the protein–nucleic acid complexes under study, which results in a higher resolution of the footprint data. Moreover, the distinct specificity in reacting only with certain functional groups like H-bond donor or acceptor groups within the different DNA grooves often enables detailed conclusions to be drawn about the mechanism of protein–DNA interaction.

For most of the chemical reagents approved for footprint analyses it must be considered that they will also react to some extent with functional groups of amino acid side chains. This demands that it must be verified during the reaction that the protein does not lose its binding activity or specificity. Hence, a gel retardation control is suggested in all cases.

Dimethyl Sulfate (DMS) A very versatile reagent approved for the analysis of many different complexes is dimethyl sulfate (DMS). DMS methylates the N7 position of guanine and the N3 position of adenine in double-stranded B-form DNA. The reagent is suitable, therefore, to map protein contacts within the major groove of DNA at G-positions, and at A-positions in the minor groove. Moreover, DMS reacts with the N1 position of adenines and to a somewhat weaker extent with N3 positions of cytosines within single-stranded DNA, which are present, for instance, in RNA polymerases–DNA complexes as a result of the melting activity during transcription. In double-stranded DNA these positions are generally not reactive because they are involved in Watson–Crick base-pairing interactions.

To identify chemical modification sites by simple sequencing gels one of the following conditions must be met: the chemical reaction must either result in a direct strand brake or the DNA chain must be destabilized by the reaction in such a way that in a second reaction a selective cleavage can be induced. Moreover, the identification of the modification site is possible if the chemical modified base causes the abortion of the DNA polymerase copying reaction during primer extension analysis. Both conditions apply for DMS, but also for many other chemical reagents. The identification of guanine methylation by DMS can also be carried out indirectly by a subsequent cleavage of the DNA chain with piperidine. The N7 methyl group, resulting from the DMS modification, confers a positive charge to the imidazole ring of the guanine base, which becomes eliminated by the piperidine reaction, followed by a strand break. To assign the DNA fragments after cleavage on denaturing sequencing gels a non-treated DNA sample is separated in parallel after a Maxam–Gilbert sequencing reaction.

In the case of a well-known sequence it often suffices to perform only the G + A Maxam–Gilbert reaction to enable a correct assignment. DMS footprint reactions can alternatively be carried out with non-labeled DNA molecules when a primer extension reaction is used for the identification of modified sites. This type of analysis takes advantage of the fact that the N7 methylated base acts as a block for the DNA-copying enzyme (e.g., Klenow fragment of DNA polymerase I). DMS footprints can therefore also be carried out with circular superhelical plasmids. Because DMS penetrates through membranes it is also possible to use the technique for *in vivo* footprint studies.

$KMnO_4$ and OsO_4 The two chemicals $KMnO_4$ and OsO_4 are able to penetrate membranes and cell walls and are therefore also suitable *in vivo* footprinting reagents. Both reagents share the same specificity but due to its lower toxicity $KMnO_4$ is used much more frequently. The reagents are known to oxidize the 5,6-double bond of thymine and with somewhat lower reactivity of cytosine within single-stranded DNA. The accessibility of the 5,6-double bonds of pyrimidines is significantly impeded in double-stranded DNA due to strong base stacking. The reaction is thus sensitive to changes in the DNA conformation, such as untwisting, but also for transitions between B- and Z-form DNA or the formation of cruciform structures. The two reagents are especially helpful for the analysis of protein binding associated with melting of the DNA double-strand or severe changes in the helix geometry. $KMnO_4$ footprints are therefore especially practical for the analysis of RNA polymerase–DNA promoter interactions, which usually result in a local melting of the DNA double-strand. Because the oxidation of the 5,6-double bond of

pyrimidines by $KMnO_4$ or OsO_4 generates a block for DNA polymerases the respective modified nucleotides can easily be identified by a primer extension analysis.

Diethyl Pyrocarbonate (DEPC) and Halo-acetaldehydes DECP and the halo-acetaldehydes, iodine- or bromo-acetaldehyde, are primarily used as probes for structural changes of the DNA conformation. Generally, these reagents do not react with B-form DNA, but they become DNA-reactive after transition into Z-DNA, the formation of cruciform structures, or melting into single-strands. DEPC carbethoxylates the N7 position of guanine and adenine as well as the NH_2 group at C6 of adenine. The halo-acetaldehydes react with non-paired cytosines, adenines, and guanines. Modified DNA strands can be cleaved with piperidine. In the case of halo-acetaldehyde modification a prior treatment with DMS (to analyze non-paired adenines or cytosines), formic acid (for non-paired cytosines), or hydrazine (for non-paired adenines and guanines) is required. This rather high level of effort has limited a broad use of this technology in recent times.

***N*-Ethyl-*N*-nitrosourea (ENU)** Apart from the reaction with some bases ENU mainly forms ethyl esters with DNA phosphate groups. It is a straightforward probe to analyze the proximity of the DNA phosphate backbone. The reaction conditions are incompatible though with most protein complexes (50 °C, 50% ethanol). The reagent is therefore mainly used for interference studies (see below). The hydroxyl radical footprint method, which likewise senses changes in the sugar-phosphate backbone, has recently largely replaced ENU modification.

Hydroxyl Radical Footprint The high resolution and the sequence-independent reaction make hydroxyl radicals, which are probably the smallest chemical reagents, a very versatile and important footprinting tool. The size of hydroxyl radicals is comparable with H_2O molecules, which enables a sequence-independent reaction with any nucleotide position of DNA- or RNA strands. Presumably, hydroxyl radicals attack the C1′ or C4′ ribose by oxidation and ultimately cause the elimination of a nucleotide leading to a strand scission. Within double-stranded DNA this reaction causes a gap (gapped duplex). The DNA ends of such gaps can either exist in 3′- or 5′-phosphorylated form. With lower probability 3′ phosphoglycolates are also formed (these different ends cause differences in the electrophoretic mobilities of fragments shorter than about 25 nucleotides). The high reactivity and the small space requirements of hydroxyl radicals give rise to a very even hydrolysis pattern over the entire length of a DNA molecule. Periodic conformational anomalies, as they occur in curved DNA or certain protein complexes, however, change the hydrolysis pattern in a characteristic way. The lifetime of hydroxyl radicals is extremely short and they only react within a small radius of their creation. Based on the very fast reaction it is supposed that the radius of diffusion is limited to about 1 nm. The observed cleavage rate at a distinct DNA position correlates with the accessibility of the corresponding bond or atom with respect to the hydroxyl radical attack. Each change in reactivity allows very precise conclusions to be drawn about the geometry of the DNA–protein complex. Hence, hydroxyl radical footprints are excellent surface probes, recording structural changes with very high sensitivity.

Hydroxyl radicals are generally formed through the Fenton reaction, in which Fe(II) ions are oxidized by H_2O_2 to Fe(III). Hydroxyl radicals are characterized by their very high reactivity and immediate attack of chemical structures in the vicinity of their creation, causing strand scissions in nucleic acids and proteins. Hydroxyl radical footprint reactions are common for the structural analysis of DNA and RNA as well as protein–nucleic acid complexes.

The generation of hydroxyl radicals is based on the Fenton reaction. In this reaction Fe(II) is oxidized by H_2O_2 to Fe(III), whereby OH^- and hydroxyl radicals $OH^\bullet$ are formed:

$$\mathrm{Fe(II) + H_2O_2 \rightarrow Fe(III) + OH^\bullet + OH^-} \qquad (32.5)$$

In the laboratory $[Fe(EDTA)]^{2-}$, hydrogen peroxide (H_2O_2), and sodium ascorbate are used for the reaction. Sodium ascorbate regenerates Fe(II) from Fe(III) in a cyclic reaction:

$$\mathrm{Fe(EDTA)^{2-} + H_2O_2 \xrightarrow{\text{Ascorbat}} Fe(EDTA)^- + OH^\bullet + OH^-} \qquad (32.6)$$

To adjust the exact reaction times the hydroxyl radical reaction can be stopped by an excess of thiourea (quenching).

Optionally, hydroxyl radical footprints are performed with end-labeled DNA. The positions are assigned by means of a Maxam–Gilbert sequencing reaction.

Because of the stereoselective reaction hydroxyl radical footprints enable a rather precise description of the topography of a protein-binding site. Hydroxyl radicals attack the

deoxyribose along the minor groove. In curved DNA the minor groove is narrowed at the inside of the curvature, which results in reduced accessibility. In the case of curved DNA it is possible therefore to recognize the exact helix periodicity, or any alteration of this parameter mediated by bound protein, in the footprint pattern. The high resolution of the hydroxyl radical reaction enables detection of even minor changes in the accessibility induced by interacting proteins. To evaluate hydroxyl radical footprints it is recommended to perform a densitometric scan of the lanes from the footprint gel. Integration of the peak areas can be used to make quantitative conclusions. Superimposing (subtraction) of the scan profiles from footprint lanes with and without bound protein will reveal architectural intricacies of the protein interacting domain. The quality of the footprints can be further improved when the DNA–protein complexes are separated from the free DNA on retardation gels after hydroxyl radical treatment. Complexes treated at single hit conditions, with a maximal one nick in the DNA, will remain intact. The DNA is extracted from the gel band and directly separated by denaturing gel electrophoresis. The background of non-bound DNA is thus eliminated, leading to a significantly improved contrast of the complex signals.

Hydroxyl radical reactions require some precautions. Radical scavengers in the solution may reduce the lifetime of the hydroxyl radicals significantly, causing a strong reduction of the band intensities. Such radical scavengers may be stabilizing substances like glycerol, which is a frequent additive in protein solutions. Consequently, glycerol-containing protein samples have to be dialyzed before the reaction. The same is true for SH-group containing solutions. Care should be taken, for instance, that the concentration of 2-mercaptoethanol does not exceed 5 mM. High concentrations of divalent cations are also known to disturb the reaction.

32.4.5 Interference Conditions

A special variant of the hydroxyl radical footprint technology is termed missing nucleotide analysis. The principle is also known as damage selection or interference footprinting.

In this type of footprint the limited hydrolysis or chemical modification of a nucleic acid is performed before it is used for complex formation with a protein. In a subsequent binding reaction only those molecules are selected that have retained their binding capability. Complexes are then separated from the free nucleic acid. Cuts or modifications at sites essential for protein binding are not present in the fraction of the complexes but are exclusively found associated with the non-bound nucleic acid. In contrast, positions identified as cleaved or modified that are only present in the fraction of the free nucleic acid are most likely involved in direct protein binding. A particular advantage of the interference footprint is that modified conditions may be applied that normally cause the dissociation of protein–nucleic acid complexes.

Because of the reactivity of hydroxyl radicals, which primarily attack the deoxyribose, this type of footprint provides exclusively information about the contacts of binding proteins with the DNA sugar-phosphate backbone. Conclusions about interactions between the bases and functional groups of the protein cannot be drawn at first. However, exactly this information is provided by the missing nucleotide principle. The method takes advantage of the hydroxyl radical cleavage chemistry to remove at random distribution single nucleotides from a DNA molecule. The reaction is performed under single hit conditions, resulting on average in less than one random gap per DNA strand. The double strands stay intact by base pairing even with a missing nucleotide in one of the strands. The DNA containing statistical nucleotide gaps is then used in a protein binding reaction and complexes formed are separated by gel retardation from the non-bound DNA. The binding experiment will select those DNA molecules in which the nucleotide gap has or has not an essential influence on the capability to form a protein complex. DNA from the complex band and the band with the free DNA is extracted and the fragments created by the prior hydroxyl radical cleavage are separated on a sequencing gel. Nucleotides that are essential for protein binding can be identified directly by comparison of the band pattern between DNA extracted from the complex and non-bound DNA. Nucleotides that are essential for binding are missing or weakly represented in the lane of the sequencing gel corresponding to the complex. In contrast, in the lane representing the free DNA strong bands can be expected at the same position (Figure 32.10).

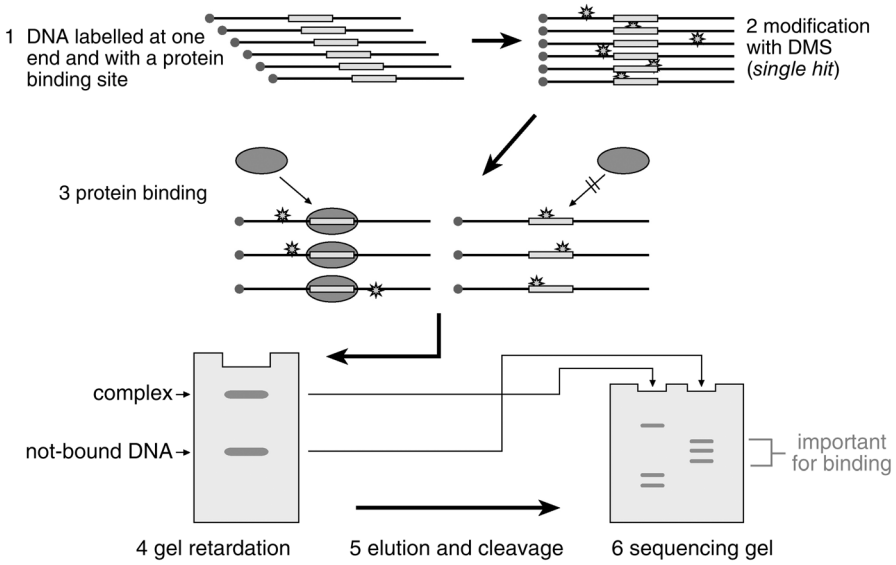


Figure 32.10 Scheme of interference footprints.

The principle of missing nucleotide analysis can also be performed with methylating reagents like DMS. In such a case one speaks of methylation interference. Because the DMS modification occurs at the nucleotide bases this method allows study of the effects of DMS-reactive positions of nucleotides on protein binding. In contrast to the hydroxyl radical reaction, methylation interference does not provide information about all bases and only purine positions can be tested. Additional reactions with alternative reagents are necessary to gain information on all DNA positions. Unlike to the missing nucleotide analysis methylation interference requires that the DNA strands must be cleaved with piperidine prior to the separation on sequencing gels.

32.4.6 Chemical Nucleases

Owing to their negative charge Fe(II)-EDTA complexes, which are used to generate hydroxyl radicals, do not bind to DNA. The generated radicals therefore react with arbitrary positions of the DNA by free diffusion. If footprint reagents are employed with the EDTA-moiety coupled to an intercalating reagent, which preferentially interacts with specific DNA structures, then a directed and less homogeneous cleavage pattern occurs. The compound methidiumpropyl-EDTA-Fe (MPE-Fe) belongs to such a group of synthetic footprinting reagents. Because the radical reaction cleaves the phosphodiester backbones of both DNA and RNA, reagents of this type are designated as chemical nucleases (Figure 32.11). MEP-Fe, like the structurally related ethidium bromide, intercalates between the stacked bases of double-stranded nucleic acids. The EDTA-iron complex is adjusted, such that it points into the minor groove. Addition of O_2 or H_2O_2 causes hydroxyl radical formation, which ultimately leads to cleavage of the phosphodiester chain. Cleavage occurs preferentially close to the minor groove. Bound proteins, which either interfere with MPE-Fe intercalation or inhibit the diffusion of hydroxyl radicals into the minor groove, cause an altered cleavage pattern. Protection from MPE-Fe-induced strand scissions therefore indicates protein binding to the minor groove. A locally reduced reactivity may on the other hand result from more complex causes. If, for instance, the DNA double-strand is unwound by protein binding MPE-Fe is unable to intercalate. Consequently, this section of DNA turns out to be protected (pseudo-protection). Because of its sensitivity towards changes in the minor groove MPE-Fe is frequently regarded as a chemical version of DNase I.

Another metal ion-chelating compound with properties of a chemical nuclease is bis(1,10-*ortho*-phenanthroline)-copper (I) complex (OP-Cu) (Figure 32.11). The reagent attaches without covalent binding to the DNA minor groove and in the presence of ascorbic acid or mercaptopropionic acid and H_2O_2 causes cleavage of the phosphodiester backbone deep within the minor groove. DNA in the B-form conformation is the preferred substrate while, under the same conditions, the DNA A-form is only weakly cleaved and DNA Z-form is not reactive at

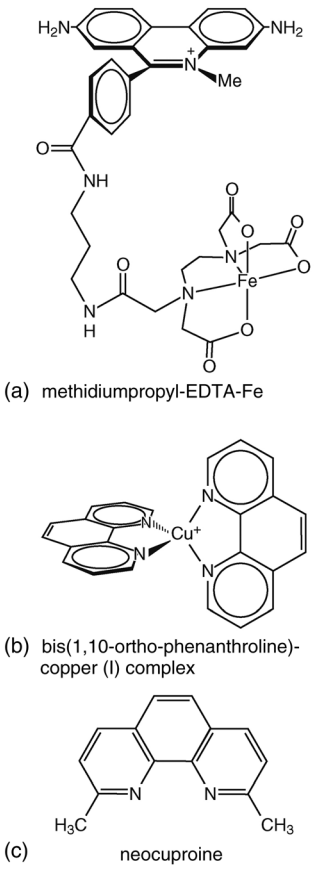


Figure 32.11 Structure of chemical nucleases.

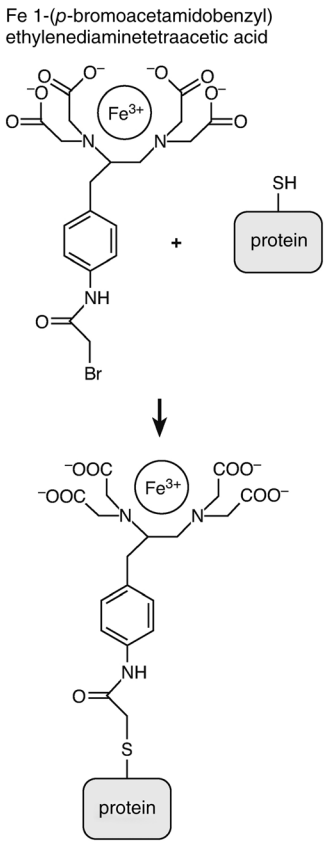


Figure 32.12 Structure of Fe-BABE, Fe 1-(*p*-bromoacetamidobenzyl)-ethylenediaminetetraacetic acid.

all. OP-Cu is much more sensitive than DNase I with respect to protein-induced conformational alterations in the minor groove. Moreover, it can cleave sequence regions characterized by frequent A:T clusters, which are only scarcely hydrolyzed by DNase I. Contrary to MEP-Fe, OP-Cu is able to cleave single-stranded regions and the fact that lower reagent concentration are required, compared to MEP-Fe, normally ensures that protein–nucleic acid interactions are not disturbed. The reagent may also be employed for structural studies of RNA–protein complexes. However, the rules that determine the cleavage efficiency are far less known due to the generally much more complex structure of RNA.

The cleavage reaction induced by OP-Cu can efficiently be stopped by the addition of 2,9-dimethylorthophenanthroline (neocuproine) (Figure 32.11). Neocuproine sequesters the available copper ions in a stable inert complex. Due to their small size the chemical nucleases described above can also conveniently be used to perform footprints directly in the retardation gel because they are able to rapidly diffuse into the matrix of polyacrylamide gels. In this case, the complex band and the band containing the free DNA may be separately subjected to a footprint analysis without prior extraction of the gel pieces. The cleaved DNA is extracted only after the treatment and directly analyzed on a sequencing gel.

A particularly elegant method for the mapping of protein–nucleic acid interacting domains is based on the targeted modification of complexes, where the interacting partners have been conjugated to reactive groups. Hydroxyl radical-generating EDTA complexes, which are conjugated to functional side chains of proteins (but also nucleic acids), have been proven to be successful. The best known of these conjugated "scissors" certainly is the iron-chelating reagent Fe-BABE. Fe 1-(*p*-bromoacetamidobenzyl)ethylenediaminetetraacetic acid (Fe-BABE), serves to generate radicals in the presence of H_2O_2 and ascorbate (Figure 32.12). The reagent can be conjugated by the bromoacetamide linker with any SH group of cysteine side chains of proteins, where it serves for the contact-dependent cleavage of a nucleic acid in complex with the protein. Contacts or close proximity of proteins and nucleic acids may thus be easily localized with Fe-BABE. The reagent can alternatively be conjugated to proteins by 2-iminothiolane and the amino group of lysine residues, which extends the application range significantly because lysines are more common amino acids in proteins than cysteines. The use of Fe-BABE-conjugated proteins does not only allow a precise localization of the protein by cleavage of the nucleic acid chain. The radicals can likewise cleave the peptide chain of neighboring proteins and thus provide information about protein–protein interactions. Because of the small activity radius of the generated radicals the cleavage positions are generally only 1.2 nm distant from the cysteine (or lysine) to which the reagent has been conjugated.

32.4.7 Genome-Wide DNA–Protein Interactions

Chromatin Immunoprecipitation, Section 31.7

Chromatin Immunoprecipitation One important method for the *in vivo* analysis of DNA–protein interacting regions is chromatin immunoprecipitation (ChIP). This technique enables us to map the genomic localization of many regulatory proteins, which is a prerequisite for the understanding of many regulatory networks. The method can be combined with DNA microarrays and "high throughput" sequencing methods and makes possible the genome-wide characterization of DNA sequences for regulatory proteins. While *in vitro* studies with purified proteins and isolated DNA fragments, along with the limitation to short DNA regions, do not reflect the actual *in vivo* situation of the DNA structure (chromatin structure or superhelicity) the ChIP methodology can provide information under almost physiological conditions. The method is furthermore not only suitable for the analysis of relatively simple bacterial DNA but also for the more complex genomes of eukaryotic organisms (Kim, T.H. and Ren, B. (2006) Genome-wide analysis of protein–DNA interactions. *Annu. Rev. Genomics Hum. Genet.*, **7**, 81–102.).

Initially, the ChIP method was based on the treatment of living cells with crosslinking reagents, usually formaldehyde, to fix existing protein–DNA complexes. The formaldehyde-crosslinked DNA–protein complexes are extracted followed by a fractionation of the chromatin and immune-affinity purification of the DNA fragments linked to the protein. This can be achieved by an antibody specific for the binding protein to be analyzed. The crosslink is subsequently reverted and the DNA binding region can be characterized by various analysis techniques like Southern blotting, PCR, or DNA sequencing methods. The original technique allows the analysis of protein-binding regions of a limited number of DNA target sequences but

no genome-wide information about the binding region of a distinct protein can be determined. The ChIP method may, however, be combined with microarray techniques, enabling the localization of protein–DNA binding sites distributed over an entire genome (ChIP-on-chip technique). It requires that the total DNA (input DNA) and the immune-precipitated DNA are labeled with two different fluorescence dyes. Both DNAs are hybridized to the same microarray chip with the input DNA serving as hybridization control. The differences in hybridization intensity on the chips are taken as a measure for the enrichment caused by the bound protein. Because the method does not require any knowledge about the potential recognition sequence it can be used for the characterization of as yet unknown binding regions leading, for instance, to a global overview of the binding sites of a distinct transcription factor on a complete genome. One important prerequisite, which poses a limitation for the method, is the availability of a highly specific antibody against the binding protein in question. If no such antibody is available, in some cases the limitation can be overcome as in organisms, which are prone for genetic manipulation. In such cases fusion of the binding protein with an epitope tag or with DNA-adenine methyltransferase for the identification of the binding positions can be helpful. The fusion of DNA-adenine methyltransferase causes a local DNA methylation in the vicinity of the protein-binding site, which can readily be identified by methylation sensitive restriction enzymes. The relatively high number of cells required for a ChIP-on-chip experiment is a further limitation. Generally, 10^7 cells are necessary for a reaction, making the method unsuitable for certain types of cells.

32.5 Physical Analysis Methods

Physical methods provide several advantages in comparison to conventional biochemical techniques such as gel electrophoretic methods combined with radioactive labeling. Firstly, most physical methods allow the observation of a binding reaction and characterization of the complexes under real equilibrium conditions, whereas gel electrophoretic analyses are performed under non-equilibrium conditions and complexes may dissociate during separation, depending on the circumstances of the electrophoresis. External conditions like temperature and salt concentration may also not be chosen freely for gel electrophoresis. A similar situation applies to enzymatic or chemical footprint methods or filter binding experiments.

32.5.1 Fluorescence Methods

As a main advantage of fluorescence-based analyses the experiments can be performed in solution without the need to separate or quantify free or complexed components. A further advantage resides in the physical principles of fluorescence itself. The timescale of fluorescence emission lies between pico- and a few hundred nanoseconds, depending on the distinct fluorophore and the condition of measurement. This timescale equals that of many molecular dynamic effects, such as molecular rotations, diffusion of small molecules, solute reorientations, movements of individual molecular domains, or the energy transfer between chromophores. Hence, fluorescence spectroscopy enables conclusions to be made about these fast processes and short-lived molecular states. Along with the analysis of equilibrium phenomena the method is of benefit for kinetic analyses in the range of milliseconds. Moreover, the sensitivity is extremely high, caused by the very low limit of detection of fluorescence dyes, and covers a range between millimolar (10^{-3} M) and approximately picomolar (10^{-12} M). Among the most frequently used analyses methods are measurements of the fluorescence anisotropy, measurements of fluorescence intensity changes, or florescence resonance energy transfer (FRET).

32.5.2 Fluorophores and Labeling Procedures

A set of different fluorescent reagents for labeling proteins and nucleic acids is commercially available. Labeling of nucleic acids is generally preferred because it is more versatile and proteins are often more difficult to obtain in pure form. Frequently used labels are for instance fluorescein- or rhodamine-tagged phosphoramidites, which are linked to the last nucleoside

position of an oligonucleotide, yielding a 5′ labeled probe. Alternatively, different coupling reagents are available in addition to spacer molecules consisting of a different number of CH_2-groups (tether). Coupling reactions may also be performed with thiol- or amino-reactive dye molecules after the nucleic acid has been modified at the 5′ end with thiol- or amino group containing phosphoramidites or ATP-γS and T4 ligase. Oligonucleotides that have been modified by amino groups may be coupled with isothiocyanate or linked to fluorescence dyes by succinimide esters.

A crucial point or the success of fluorescence-based binding studies is the purity of the fluorescence probe. It is essential for the sensitivity of the respective study that 100% of the DNA or RNA are labeled with the fluorescence dye. Any contamination of free fluorophore should also be avoided because it will contribute to the anisotropy of the overall signal, notably limiting the sensitivity of the study. HPLC-purification of the oligonucleotides after labeling is the method of choice. Oligonucleotide purification kits, size exclusion chromatography (e.g., Sephadex G-10), or gel electrophoresis are also an option. Preparative gel electrophoresis under denaturing conditions has proven of value as a universal and very efficient purification method. For the preparation of fluorescence labeled DNA double-strands the labeled single-strands are usually hybridized followed by the separation of the fluorescence labeled non-base-paired single-strands by native gel electrophoresis.

32.5.3 Fluorescence Resonance Energy Transfer (FRET)

FRET, Section 16.7.1–16.7.4

A special application of fluorescence analysis for the characterization of spatial relationships of complex macromolecules is the fluorescence resonance energy transfer (FRET) technique. The method is particularly suitable for the determination of structural changes induced by protein–nucleic acid interactions. FRET allows measurement of the distances between fluorescent dyes residing at defined positions of a macromolecule, thereby providing three-dimensional structural information, which is otherwise very difficult to obtain. The method can successfully be used, for instance, to determine in a qualitative and quantitative way the protein-induced DNA-bending of labeled molecules. For the analysis, each donor and acceptor fluorescence dye must be linked at the ends of a DNA fragment containing a protein binding site. Of course, the end-to-end distance should not exceed the range limit of the fluorescence energy transfer (Förster transfer).

Fluorescence resonance energy transfer describes a non-radiative transfer of energy (without emission of a photon) from the excited state of a chromophore, the energy donor D, by intermolecular long-range dipole–dipole coupling to a neighboring chromophore, the energy acceptor A. The efficiency of the transfer is inversely proportional to the sixth power of the distance between the fluorescence dyes. Hence, this limits the method to measurements between 1 and 10 nm. Quite obviously, for a FRET system to be efficient the fluorescence spectra of D and the absorbance spectra of A must be sufficiently overlapping. A further critical parameter for FRET analysis resides in the orientation of the two dyes with respect to each other. The transfer efficiency depends on the relative orientation of the diploes of the donor and acceptor. To enable highly efficient transfer the transition dipoles of donor and acceptor should preferentially be parallel. The relative orientation of the dipoles with respect to each other is given by the orientation factor κ^2, which is not precisely known in most cases and the simple assumption that both chromophores can rotate freely and fast with respect to the lifetime of the excited state of the donor ($\kappa^2 = {}^2/_3$) is certainly not always correct. If one considers, what might be true for many donor–acceptor pairs, that one chromophore is static while the other rotates fast and freely then the error in the distance determination R_0 adds up to simply 12%. The distance at which the energy transfer is 50% is defined as the Förster distance (R_0). The magnitude of R_0 depends on the spectral properties of the donor and acceptor molecules and is calculated by:

$$R_0 = 9.79 \times 10^3 \left(JQn^{-4}\kappa^2\right)^{\frac{1}{6}} \tag{32.7}$$

where: J = the spectral overlap integral,

Q = the fluorescence quantum yield of the donor in the absence of the acceptor,

n = the refractive index of the medium,

κ^2 = the dipole orientation factor.

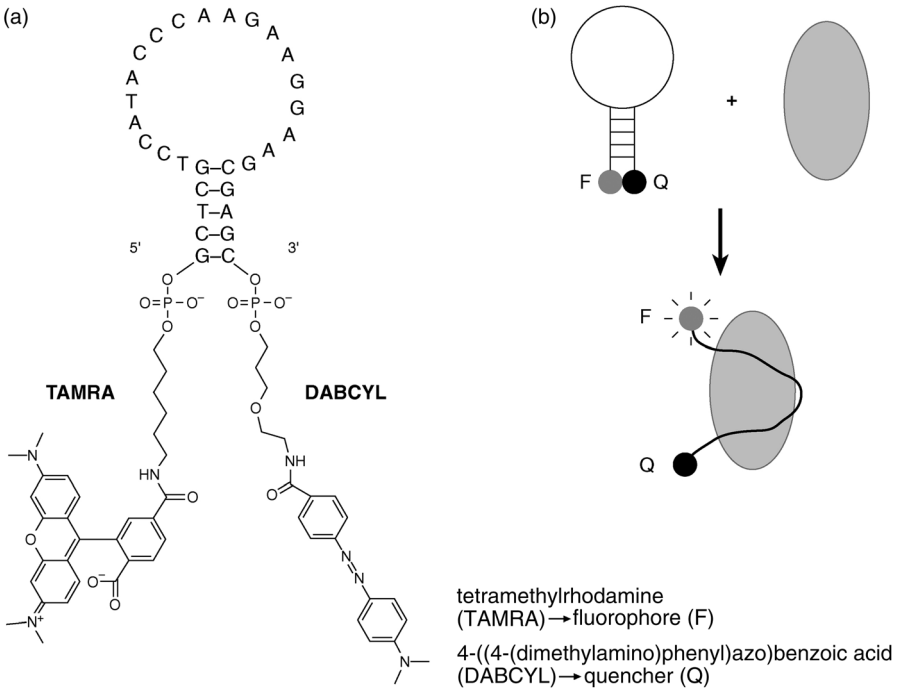


Figure 32.13 Molecular beacons for the detection of protein–nucleic acid complexes. (a) Structure of a molecular beacon; due to the close proximity of the fluorophore (F) to a quencher (Q) no fluorescence signal is emitted. (b) The interaction of a protein causes the spatial separation of fluorophore and quencher, which results in the emission of a fluorescence signal. Source: taken from Li, J. J., Fang, X., Schuster, S. M., Tan, W. (2000) *Angew. Chem., Int. Ed.*, **39**, 1049–1052. With permission, Copyright © 2000 Wiley-VCH Verlag GmbH.

It is known that many dyes, which are used for labeling of nucleic acids, interact with neighboring bases. This can significantly contribute to quenching of the excited state of the chromophore. Moreover, since these interactions are sequence-dependent the newly formed nucleic acid–dye complexes may have completely different spectroscopic and stereochemical properties. To enable identical FRET efficiencies for quantitative comparisons it is important, therefore, that all labeled probes have the same nucleotide sequence in the vicinity of the fluorophore.

32.5.4 Molecular Beacons

Molecular Beacon, Section 28.1.3

The descriptive name molecular beacons denotes a special application of fluorescence analyses that depends on changes of the fluorescence of a chromophore by a quencher molecule located nearby. Molecular beacons consist of a hairpin nucleic acid with a fluorescence molecule linked to one end and a quencher molecule to the other end. The spatial proximity of the quencher to the fluorophore prevents the emission of a fluorescence signal. When the hairpin structure is disrupted by the interaction with other biomolecules (DNA, RNA, or proteins) the distance increases between fluorophore and quencher, which causes an increase in the fluorescence intensity (Figure 32.13). Molecular beacons are therefore favorable tools for the identification of protein–nucleic acid interactions. The principle can also be applied in intact cells. In such cases the probes can be introduced into the cells through liposomes.

32.5.5 Surface Plasmon Resonance (SPR)

Surface Plasmon Resonance (SPR), Section 16.6

Surface plasmon resonance (SPR), often designated according to the company logo *Biacore,* describes an optical method measuring changes of the diffraction index close to a sensor layer (~300 nm) between glass and metal. The surface forms the bottom of a tiny cell (20–60 nl) through which an aqueous solution containing a binding partner (analyte) flows in a continuous way. Onto the sensor surface (usually a chip with a thin gold film) a binding protein (ligand) is immobilized. Binding of the analyte to the ligand increases the molecular mass at the surface, resulting in an increase in the refractive index. Several chemical procedures to immobilize the ligand on to the chip surface are available. The choice of the coupling conditions depends on the

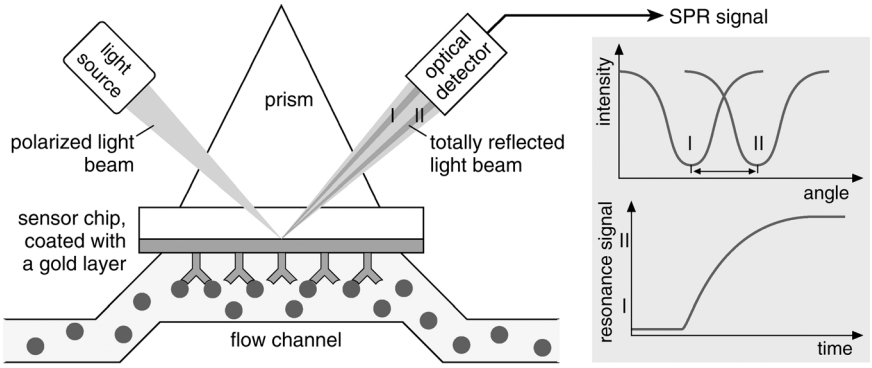


Figure 32.14 Principle of an SPR measuring device.

chemical properties and expected sizes and structure of the complexes between the ligand and analyte.

The physical principle of surface plasmon resonance is relatively complex and takes place when a monochromatic planar polarized light beam is totally reflected at the interface of a glass surface, which is coated on its outside with a gold layer. This creates two different optical waves: one wave with exponentially decreasing intensity (evanescent) resulting from total internal reflection, and one wave propagating through the metal. At a certain angle the interaction between both waves causes a drop of the surface-reflected light (plasmon resonance). This drop can be recorded. The resonance conditions depend on the material that is adsorbed to the metal layer (analyte). The resonance energy shows an almost linear dependence on the mass concentration of biomolecules, like proteins, DNA, or RNA, that are fixed to the sensor surface. The SPR signal, recorded in resonance units (RU), is a measure of the concentration of the mass on the surface of the sensor chip (~1000 RU corresponds to the adsorption of 1 ng mm^{-2} protein or nucleic acid; this results in a deflection of the reflection angle by 0.1°). SPR measurements allow us to record the association and dissociation of analyte and ligand and to determine equilibrium as well as rate constants of the interaction (Figure 32.14).

32.5.6 Scanning Force Microscopy (SFM)

Atomic Force Microscopy, Section 20.2

Scanning force microscopy (SFM), also sometimes termed atomic force microscopy (AFM), is a typical method for studying single molecules. With SFM, surfaces are scanned in the nanometer range with a very fine needle fixed to a cantilever. To exactly maneuver the needle over the sample piezo-electric elements are used. The needle hovers over the sample surface, held in place by a minute spring force. In a line-by-line scanning of the surface structure the cantilever tip is deflected. This deflection is recorded by a laser, which is reflected from the backside of the tip. This information is recorded by a photodetector and converted into a three-dimensional image. SFM can routinely visualize structures in the atomic range (between 0.1 and 10 nm). Measurements are possible in air or in aqueous solutions and, hence, at physiological conditions. There are different ways the measurements can be performed. In the contact mode the tip of the needle remains at very close distance to the sample material to be scanned. Thereby, the tip interacts with the sample by van der Waals, capillary, or electrostatic forces. As a disadvantage, shearing forces may distort or destroy the sample material. Alternatively, measurements can be performed by the tapping mode. Thereby, the spring that fixes the tip oscillates at high frequency. The measuring procedure is largely without contact. If the fine tip touches the sample the oscillation amplitude goes down. A regulatory circle keeps the amplitude constant so that only very light contacts occur. As an advantage the lateral resolution is higher with lower application forces, which preserves the sample. For the measurements the sample, purified as highly as possible, is deposited in a dilute buffer onto a freshly cleaved mica surface. Adsorption to the mica can be improved by Mg^{2+} ions. Preferred objects for SFM are primarily protein–nucleic acid complexes of higher molecular weight, such as DNA bound to RNA polymerase or chromatin complexes. However, successful analyses have been performed with small proteins like the transcription factors Cro or FIS with molecular weights of 2×7.6 kDa and 2×11.2 kDa, respectively. The method is especially capable of

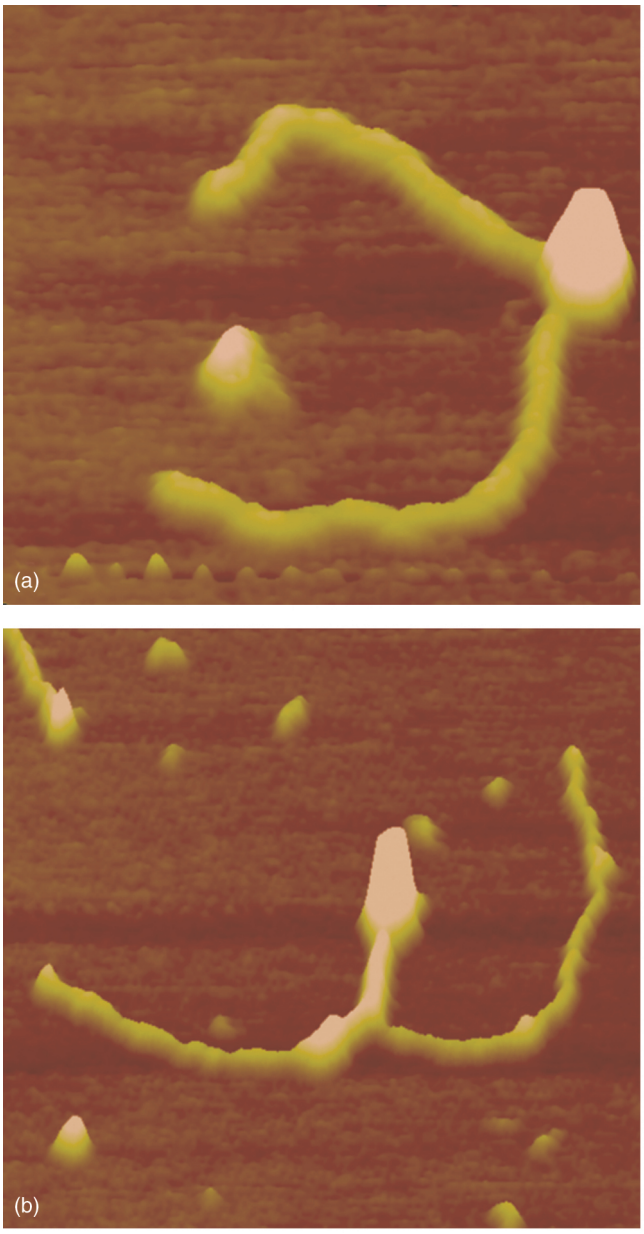


Figure 32.15 Analysis of the binding of RNA polymerase to a promoter DNA by scanning force microscopy (SFM). (a) RNA polymerase in complex with a DNA fragment containing a bacterial rRNA promoter. (b) The same complex in the presence of the repressor protein H-NS. The DNA is fixed by H-NS around the RNA polymerase, which prevents it from leaving the promoter. Source: courtesy of Dr. R. Dahme, Leiden, The Netherlands.

detecting conformational changes of macromolecules, enabling the visualization of DNA bending, loop formation, or topological alterations of circular molecules (Figure 32.15).

32.5.7 Optical Tweezers

Optical tweezers use light to manipulate microscopically small objects, down to the size of an atom. The method takes advantage of the fact that the focus of a laser beam can fix a small particle (optical trap). With such a device forces in the range of pico-Newtons (pN) and deflections in the range of nanometers can be precisely determined. If, for instance, the interacting forces of a molecular system should be determined one component of this system is bound to a solid support while a second component (often fixed to a silica bead) is kept in an optical trap. The deflection of the silica bead within the focused laser beam can be measured by the restoring force. With such a device it is possible to determine the interacting forces between macromolecules or, for instance, the motion and the forces involved in an RNA polymerase synthesizing an RNA chain while traveling along a DNA molecule (Figure 32.16).

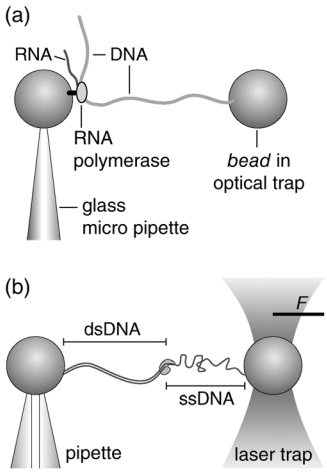


Figure 32.16 Examples of the analysis with optical tweezers. (a) Schematic depiction of an experiment to measure the force of a transcribing *Escherichia coli* RNA polymerase. (b) Schematic arrangement to measure the forces of the transcription and exonuclease reaction of T7 DNA polymerase. Source: taken from Wuite, G.J. *et al*. (2000) *Nature*, **404**, 103–106. With permission, Copyright © 2000, Rights Managed by Nature Publishing Group.

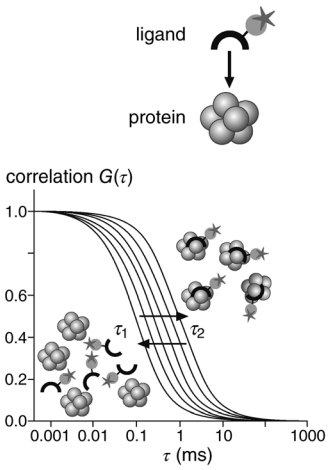


Figure 32.17 Example of an FCS measurement. The change of the diffusion time is recorded that results from the formation of a complex between a large molecule (protein) and a small fluorescence labeled oligonucleotide (ligand).

32.5.8 Fluorescence Correlation Spectroscopy (FCS)

Fluorescence correlation spectroscopy (FCS) is a very fast and dynamic method for the analysis of DNA– and RNA–protein interactions in solution. The method combines laser technology and confocal microscopy. Although rather long standing the method has been further developed to become a promising technique by introduction of modern laser microscopy. In FCS the random motion of fluorescence labeled molecules is recorded in a very small volume, which roughly corresponds in size to the volume of an *Escherichia coli* cell (10^{-15} l, 1 femto-liter). A focused laser beam irradiates this small volume. The application is based on fluctuations of the fluorescence signal resulting from single molecules, which diffuse around the solution studied. The diffusion times of the molecule can be obtained from the fluctuation data, which directly correlates with the mass of the particles. Every change of the molecular mass, like complex formation with another macromolecule, strongly alters the diffusion times and can be used for the analysis of thermodynamic or kinetic parameters of the interacting partners. In this way mobility coefficients and characteristic rate constants of inter- or intramolecular reactions can be determined at nanomolar concentrations. The time-resolution of the method spans from a millisecond to more than 10 s. The measurement requires, along with a fluorescence correlation spectrophotometer, a fluorescence labeled probe. Relatively small fluorescence labeled ligands that bind to heavy (about a factor of 8–10 bigger) non-fluorescent molecules with large diffusion coefficients are ideally suited for the analysis (Figure 32.17). Lately, also devices are at hand for the simultaneous analysis of two differently fluorescence labeled components (two-color detection).

32.6 RNA–Protein Interactions

Remarkably, RNA molecules within a cell rarely exist as single molecules. At almost all times they are complexed with proteins and execute their functions as ribonucleoprotein (RNP) particles. RNPs can be very complex, like ribosomes or spliceosomes. However, there are also simple binary complexes, like RNase P or telomerase. Often, RNA molecules, depending on their functional tasks in the cell, may rapidly change between different complexes. If tRNA is taken as an example it almost never exists without a protein partner in the cell. Its functional path right after synthesis and maturation is accompanied with aminoacyl synthetase, which catalyzes the link with the specific amino acid. Following amino acylation tRNAs form a ternary complex with the elongation factor EF-Tu•GTP. They are then docked to the ribosomal A-site where they change to the P- and E-site before they are amino-acylated again after leaving the ribosome. Clearly, such RNA molecules must possess a complex pattern of different specific interaction mechanisms.

32.6.1 Functional Diversity of RNA

In addition to the general task as mRNAs, for information transfer from DNA to protein, RNA molecules have many different functions. For instance, tRNAs, a family of stable RNAs, function among several alternative tasks as adaptors for amino acids, which they activate and transport to the protein synthesis machinery. A different set of RNA molecules, ribosomal RNAs (rRNAs), not only constitute the most important structural part of all ribosomes but are also active in binding and catalytic steps during protein synthesis. Generally, RNA molecules exerting catalytic functions are collectively termed ribozymes. Among the catalytic functions of natural ribozymes are for instance nucleolytic activities (hydrolysis of nucleic acids), formation of phosphodiester bonds (ligation or splicing), or peptide bond formation. In many higher organisms guide RNAs (gRNAs) play a role in RNA editing. Moreover, small nuclear RNAs (snRNAs) or small nucleolar RNAs (snoRNAs), which are responsible for the splicing process or the maturation and sequence-directed modification of eukaryotic rRNAs, respectively, are of special importance. Recently, in all organisms studied so far, many small non-coding RNAs (ncRNAs) have been characterized, which have variable regulatory functions at the different levels of gene expression. Among those RNAs are the small interfering RNAs (siRNAs) responsible for the RNA interference phenomenon or micro RNAs (miRNAs). Moreover, the

recently discovered small RNAs (crRNAs) involved in the prokaryotic defense system CRISPR against foreign DNA belong to the group of versatile regulatory RNAs.

32.6.2 RNA Secondary Structure Parameters and unusual Base Pairs

What are the reasons for the specific differences in the interactions of proteins with either DNA or RNA? Despite the primarily rather small chemical differences between the two classes of molecules, DNA and RNA, namely the change of 2′-deoxyribose to ribose and a swap of a methyl group in thymine against a hydrogen atom in uracil, there are enormous differences in the structures and function of the two macromolecules. The preferred sugar pucker in 2′-deoxyribose is the C2′-endo form while in ribose, due to the larger space requirements and the ability of the C2′-OH group to form H-bonds, the C3′-endo form is prevalent. As a consequence of this conformational difference the distance to neighboring phosphates is shortened, which in double-stranded polynucleotides causes a stabilization of the A-form helix. In the A-form helix, in contrast to the B-form helix, the standard conformation of DNA, the base-pairs are not oriented at right angles with respect to the helix axis but are tilted and shifted slightly lateral out of the helix center (slide, shift). This causes a change in the dimensions of the major and the minor grooves compared to the B-form helix. The major groove changes to a deep narrow groove in which the functional groups of the bases are not accessible for protein contacts as in double-stranded B-form DNA. The minor groove of the B-form is converted into a shallow groove in the A-form. Contrary to the deep major groove, the minor shallow groove of the A-RNA helix is easily accessible for functional groups of proteins. However, the known structures of natural RNA that have been solved until now indicate that helical sections are rarely longer than a complete helical turn without being interrupted by mismatches, bulges, or other secondary structural elements, which enables access to the deep groove from the helix ends. The discontinuation of the regular helix structure enables a multiplicity of structural variants, which deviate considerably from the A-form, producing an inexhaustible repertoire of options for interactions (Figure 32.18).

Additional structural peculiarities of RNA molecules are deviations from the standard Watson–Crick base pairs, like the existence of wobble, Hoogsteen, or reversed-Hoogsteen pairs, mismatch pairs, bulges, stem-loop structures, bifurcations, tetraloops, or tertiary interactions resulting from base-triples or pseudoknots (Figure 32.19). This structural multiplicity makes up for the somewhat lower information content of the RNA helix grooves, which, as indicated above, are anyway rather rare in natural RNAs without irregularities. RNA-binding proteins, therefore, recognize only in special cases perfect double-helical elements with a conserved sequence. In contrast, single-stranded regions in between secondary structure elements are generally recognized in which the functional groups of the bases are exposed for sequence-dependent interactions. Additionally, intermolecular stacking interactions, which are normally very scarce in DNA, are often found in RNA–protein complexes. The energetic contribution to the stability of the complexes is especially high ($\sim 3\,\mathrm{kcal\,mol^{-1}}$). For the description and a molecular understanding of RNA–protein interactions exact knowledge of the complex structure folding of RNA molecules is important.

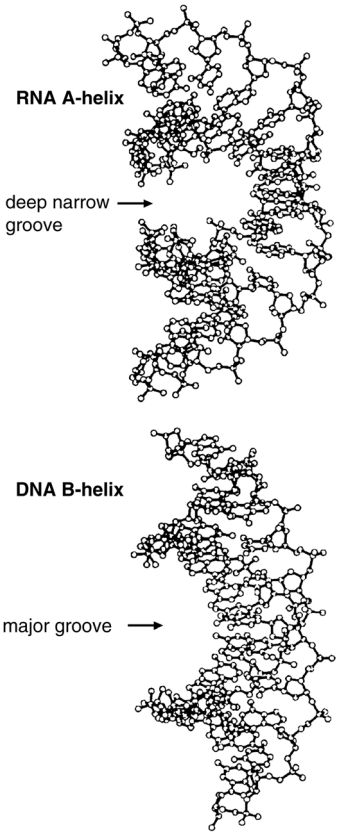


Figure 32.18 Differences in the geometry of the helical grooves between DNA and RNA. The accessibility of reactive groups of proteins is strongly restricted in the deep groove of the RNA helix.

32.6.3 Dynamics of RNA–Protein Interactions

The specificity of DNA–protein interactions depends to a large degree on distinct interactions of amino acid side chains with functional groups of bases from the DNA strands in the major or minor grooves. The specificity of the interaction is often based on symmetrical DNA-structure elements (palindromes, sequence repetitions). Additional interactions with the sugar-phosphate backbone enhance the specificity of the complex formation. In other words, the interactions are predominantly based on the helical nature of the nucleic acid chain. In contrast, the specificity of RNA–protein interactions, in agreement with the much higher variability of the RNA 3D-structure, is considerably versatile. Moreover, RNA molecules are (often) much more flexible, which is a prerequisite for many catalytic processes. A clear tendency of RNA-binding proteins that can be noted is to “freeze” the correct structure from an ensemble of alternative structures.

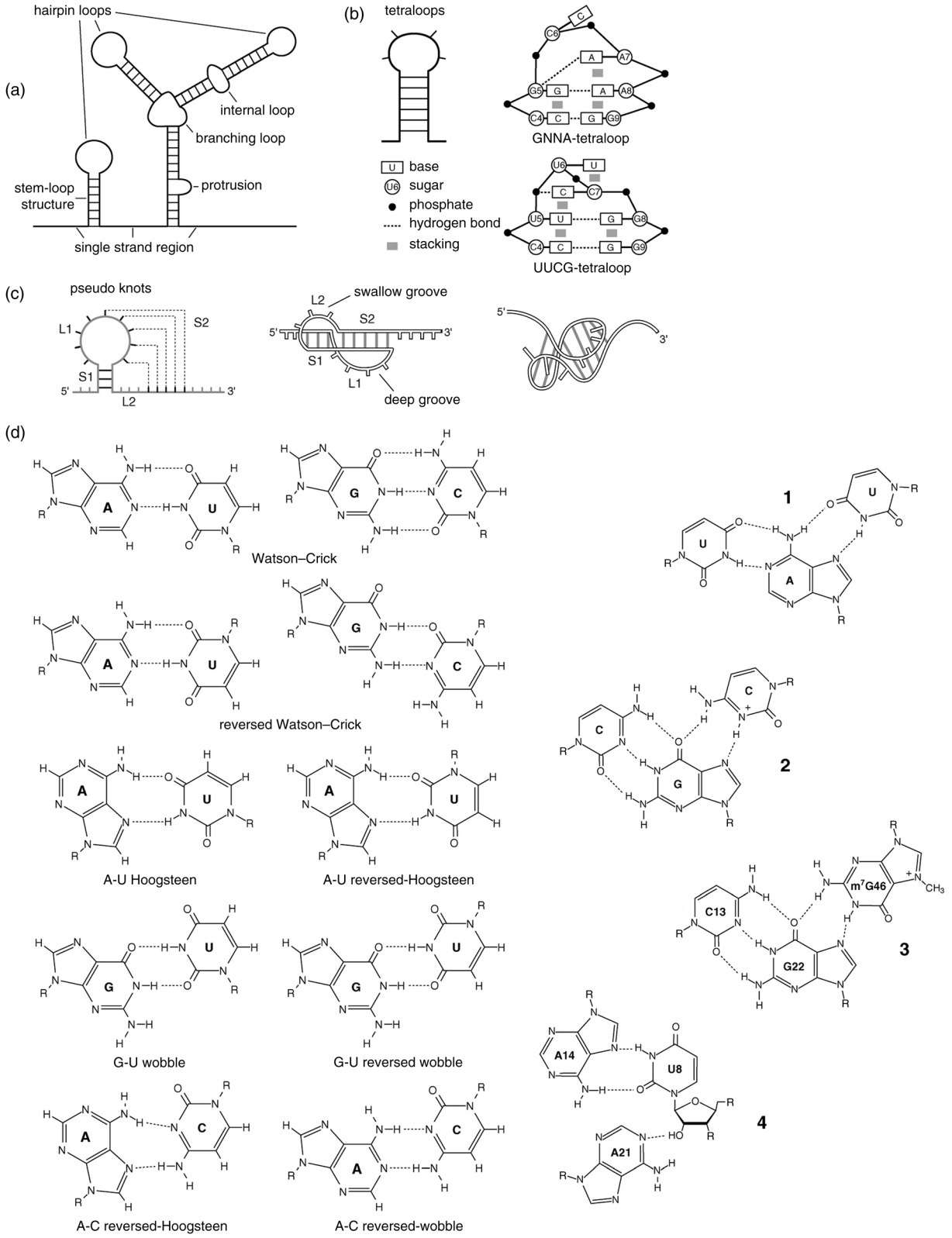


Figure 32.19 Characteristic structural elements of RNA. (a) Different secondary structures; (b) examples of special secondary structures; (c) a pseudoknot as an example of a tertiary structure; (d) examples for special base-pairs and base-triples: (1) Hoogsteen-base-triple A:U:U, (2) protonated Hoogsteen-base-triple C^+:C:G, (3) purine-base-triple in $tRNA^{Phe}$, and (4) adenosine-H-bound to a 2′-OH-group of a reverse Hoogsteen-A:U pair in $tRNA^{Phe}$. (Batey, R.R., Rambo R.P., and Doudna, J.A. (1999), Tertiary Motifs in RNA Structure and Folding. *Angew. Chemie Int. Ed.* **38**, 2326–2343)

Presumably, conformational adaptation during RNA–protein recognition plays a much greater role than for the interaction between DNA and protein.

Surprisingly, the binding energies of RNA–protein complexes do not correlate with the size of the interaction surface, which is predominantly participating in van der Waals interactions and makes an entropic contribution to the replacement of solvent molecules and ions. For example, the interface of the very stable U1A protein–RNA complex ($K_D \approx 10^{-11}$ M) is very small and the binding includes an entropically expensive step, which converts a disordered single-strand RNA structure into an ordered secondary structure. It is assumed that RNA–protein interface regions are not completely rigid. Rather it is supposed that a compromise between rigidity and flexibility exists, which determines whether either high specificity, with the necessary very high entropic expenses, or the absence of any selectivity exists. An important parameter determining the binding specificity resides in the energetics necessary for the intrinsic RNA folding. For some RNA–protein complexes, melting of the RNA secondary structure causes an up to 10^5-fold lower binding energy. If one assumes that a preformed RNA secondary structure reflects exactly the spatial counterpart of a rigid protein binding domain then the entropic expenses for RNA folding, when bound to the protein surface, are reduced by the fraction of the electrostatic binding energy contribution, which derives from the interaction with the phosphate backbone. If one and the same protein recognizes two different RNA molecules the one that more readily forms its correct structure will be bound much better, no matter if all contacts between the protein and the two RNAs are identical in the bound state. The formation of RNA–protein complexes is therefore clearly a highly dynamic process. The recognition directed by the RNA structure generally changes the RNA conformation to form a fitting intermolecular binding surface. Hence, knowledge of the static RNA structure is not sufficient to understand the specificity of RNA–protein interactions on a quantitative level and it is necessary that the information is complemented by additional thermodynamic and kinetic studies.

32.7 Characteristic RNA-Binding Motifs

Commonly, RNA-binding proteins consist of independently folded compact globular αβ-recognition domains, which may occur singly or in multiple repetitions. Obviously, αβ-structures provide a particularly suitable platform enabling specific interaction with RNA molecules. The most prominent αβ-domains of RNA-binding proteins are the ribonucleoprotein (RNP) domain, the K-homology domain (KH), and the double-stranded RNA binding domain (dsRBD). The arrangement of the αβ-secondary structure motifs between these three RNA-recognition modules is generally different. The RNP domain is characterized by a repetitive βαβ-arrangement, dsRBD proteins have an αβββα organization, and KH proteins exhibit a βααββα secondary structure fold.

Another class of RNA-binding proteins is characterized by a basic domain. This domain consists of a stretch of about 10–15 amino acids with a very high content of arginines and lysines (arginine-rich motif, ARM). In contrast to the recognition motifs described above ARM proteins do not have independently folded compact domains. Rather, the structure of the binding domain adapts itself to the respective RNA secondary structure. Moreover, the basic amino acids are not found at conserved positions but occur as clusters with relatively high variability. RNA secondary structures that interact with ARM proteins often show structural irregularities like bulges or internal loops within a regular RNA double helix. The major groove of the RNA helix is widened such that a β-sheet or a α-helical secondary structure can enter, allowing specific interactions with functional groups of the RNA bases from the major groove. The specificity depends to a great extent on how exactly an adaptation to the target RNA can occur. This becomes apparent by the observation that short synthetic arginine-rich peptides already possess notable affinities to RNA molecules (K_D values in the range of 10^{-9} M). According to thorough studies those peptides do not differentiate by more than a factor of two between related and non-related RNA molecules. In contrast, a specific ARM-RNA-binding protein like (HIV-1) Rev protein binds a specific RNA 1000-fold better than a non-related RNA.

The RBD domain has already been described as a RNA-binding motif rather early on. It is known to occur for instance in the protein U1A and the U1-snRNA. Proteins from this family exist in very different organisms and occur in all possible compartments and organelles.

Individual proteins from this family bind to various different RNA molecules. Based on their large number and the large functional diversity, proteins with RBD motifs belong to the biggest family of RNA-binding proteins. Usually a RNA hairpin is recognized by a βαβ-βαβ-protein fold with characteristic aromatic side chains. Examples for RBD proteins are hnRNP A1, PABP, U2AF (U2 auxiliary factor), or general splicing factor SF2/ASF.

A relatively common motif found in RNA-binding proteins is the KH-domain. The name is derived from hnRNP K, a human RNA-binding protein. The KH motif exists singly or in several copies and confers binding properties towards single-stranded RNA. The domain is composed of approximately 50 amino acids containing a conserved octapeptide, -Ile-Gly-X_2-Gly-X_2-Ile- (X denoting any amino acid). Structure predictions indicate a three-stranded β-sheet with two α-helices arranged opposite. The ribosomal protein S3 is a typical example of a protein with KH domain. The occurrence of KH domains in very different organisms suggests that the KH motif is in evolutionary terms a rather ancient protein motif.

Another well-defined RNA-binding domain consists of the so-called RGG box. This motif is built from 20 to 25 amino acids with tightly staggered Arg-Gly-Gly (RGG) sequences often separated by aromatic amino acids. The number of RGG motifs (RGG boxes) within different proteins can vary between 6 and 18. The arginine residues within the RGG boxes are often modified as N^G,N^G-dimethyl-arginine, which contributes to the variation of the steric properties of the motif. RGG motifs are often found in combination with other RNA-binding domains enhancing the specificity.

The double-stranded RNA-binding domain (dsRBD) is an approximately 70 amino acid long sequence region, intermingled with numerous basic amino acids, recognizing double-stranded RNA. Proteins containing dsRBD motifs do not bind to double-stranded DNA, however, which indicates that the motif specifically recognizes the geometry of the A-form helix. Often, dsRBD motifs exist in multiple copies.

Rather scarce as RNA-binding motifs are zinc finger domains although some exhibit additional DNA binding ability. TFIIIA, for instance, binds with a nine-fold zinc finger both the 5S rRNA gene and to the 5S rRNA itself, whereby the three middle zinc finger domains are predominantly responsible for the RNA binding. Similar properties for the interaction with DNA and RNA are known for Y box proteins, which comprise several eukaryotic transcription factors. These proteins can additionally recognize RNA and often function to mask mRNAs. With respect to their structure they resemble bacterial cold-shock proteins (CSPs).

Numerous RNA-binding proteins are characterized by a so-called oligonucleotide/oligosaccharide binding fold (OB fold). This motif consists of a five-stranded β-barrel structure. Members of the class OB fold proteins are, next to tRNA synthetases, the bacterial termination factor Rho, which recognizes C-rich RNA, or the *trp* RNA-binding attenuation protein (TRAP). For both proteins the structure has been solved at high resolution. The Rho factor binds as a hexameric protein to C-rich RNA with little or no secondary structure. In contrast, TRAP recognizes a specific sequence within the *trp* operon RNA-leader. Binding results in the distortion of a hairpin structure, which normally acts as a terminator structure for RNA polymerase. The protein is composed of 11 identical subunits consisting exclusively of β-sheet structures. The recognition is carried out by 11 repetitive trinucleotide sequence repeats (G/UAG) separated by each two nucleotides. The affinity between TRAP and the leader RNA is modulated by tryptophan. A stable RNA complex will only form when all TRAP subunits contain bound L-tryptophan. Both the termination factor Rho and the RNA-binding protein TRAP are impressive examples that binding of a protein to its target RNA is accompanied by dramatic changes of the RNA secondary structure with distinct functional consequences (riboswitch).

32.8 Special Methods for the Analysis of RNA–Protein Complexes

Many of the analysis techniques described for DNA–protein complexes are suitable for RNA molecules as well. The same applies for the physical methods, outlined in Section 32.5, which are generally suitable for studies of both types of complexes. The following section summarizes those types of analyses that have especially been developed for the handling of RNA–protein complexes taking advantage of RNA-specific enzymes or chemicals.

Table 32.1 Specificities of RNases.

RNase	Substrate	Cleavage
RNase T1	Single-stranded GpN	Gp (3′-P)
RNase U2	Single-stranded ApN	Ap (3′-P)
RNase CL3	Single-stranded CpN ≫ ApN > UpN	Cp, Ap, Up (3′-p)
RNase T2	Single-stranded NpN	Np (3′-p)
Nuclease S1	Single-stranded NpN	pN (5′-p)
RNase CVE (V1)	Double-stranded NpN	pN (5′-p)

32.8.1 Limited Enzymatic Hydrolyses

For the analysis of RNA structures several specific enzymes (RNases) are available, which differ notably in their specificity and structure selectivity from DNases (Table 32.1). Although RNases can be divided in endo- or exonucleases no parallel exists to the large number of DNA sequence-specific endonucleolytic hydrolyzing restriction enzymes. Instead, several RNases are known that are highly specific for a single nucleotide and that hydrolyze the phosphodiester chain in a way that the phosphate remains either at the 3′- or the 5′-site of the recognized nucleotide. Because, usually, those functional groups of the bases are recognized, which are generally involved in Watson–Crick-type base pairing, and hence not freely accessible in the paired state, most RNases are preferentially single-strand-specific. There are also RNases without distinct base specificity, which are capable of hydrolyzing the RNA chain after each possible nucleotide A, U, G, and C with statistic frequency. By combination of different RNases it is possible to gain information about each nucleotide of an isolated RNA molecule as well as within a protein complex. In addition to the base-specific RNases some enzymes are known with a distinct specificity for RNA secondary structure, which only cut either single-strand or double-strand RNA. The use of these enzymes represents the experimentally simplest and generally applicable way to solve the secondary structure of RNA molecules, single or in complex with a protein.

32.8.2 Labeling Methods

The analysis following enzymatic hydrolysis is basically always possible by direct gel electrophoresis of the resulting cleavage products of radiolabeled RNA molecules. The labeling can be done either at the 3′- or at the 5′ ends. For the direct analysis of end-labeled RNA there is, however, a certain length limitation (about 200–400 nucleotides) depending on the resolution power of the sequencing gel system. Hence, information will only be accessible for nucleotides that are not further away than 200–400 nucleotides from the labeled end. Labeling of the 5′-ends is performed as described for DNA fragments by the transfer of a [^{32}P]-phosphate group from γ-[^{32}P]-ATP catalyzed by T4-polynucleotide kinase. In the case of already 5′-phosphorylated RNA fragments the 5′-phosphate is removed by alkaline phosphatase or a polynucleotide kinase exchange reaction with γ-[^{32}P]-ATP is performed. Labeling at the RNA 3′-ends is carried out by ligation of a short radioactive labeled nucleotide diphosphate (typically 5′-[^{32}P]-pCp) to the free 3′-OH-group of the RNA. The appropriate enzyme is T4-RNA ligase. To facilitate the assignment of bands on the denaturing sequencing gels it is common to co-separate an aliquot of the labeled RNA after a mild statistical alkali hydrolysis. This generates a “ladder” representing cleavage products at all nucleotide positions. In addition, a limited hydrolysis of the labeled RNA with RNase T1 has proven to be of value due to the high specificity, which indicates the positions of all guanines within the RNA. The following should be considered for the analyses. With the enzymatic cleavage of RNA molecules it is possible that the first cut already induces dramatic structural changes, which may cause the native complexes to fall apart. As a consequence secondary cuts may appear rapidly, which falsifies the analysis results. There is,

however, an elegant way to distinguish such secondary from primary cuts. It takes advantage of the fact that primary cuts should be visible as well with the 5′ end-labeled as with the 3′ end-labeled RNA molecule while secondary cuts will only show up in one of the labeled samples. For a correct analysis it is reasonable, therefore, to use samples that are labeled at either end and to interpret only those results that are evident with both labeled samples.

32.8.3 Primer Extension Analysis of RNA

A second method used to assign cleavage or modification sites within an RNA molecule is the primer extension reaction, which can in principle be used for the analysis of protein–DNA complexes. Yet, primer extension reactions with RNA molecules are not performed with a DNA polymerase but with reverse transcriptase. This enzyme, starting from a complementary DNA oligonucleotide hybridized to the RNA, synthesizes a cDNA in the presence of the four dNTPs as substrate. In this reaction the RNA sequence is copied from the 3′ position of the annealed oligonucleotide into the 5′ direction. The primer is extended until the 5′-end of the RNA is reached. This means that all newly generated 5′-ends caused by the cleavage reaction result in a new abortive primer extension product on the denaturing gels. Furthermore, the reverse transcription reaction stops the primer extension at bases modified at positions that interfere with the Watson–Crick base-pairing. These are modifications at N1-Ade, N1-Gua, N2-Gua, N3-Cyt, and N3-Ura. The N7-Ade modification, introduced by the DEPC reaction (see below), also suffices for a primer extension abort although this position is not involved in typical Watson–Crick base-pairing. Modifications at N7-Gua require a subsequent strand break reaction to be detected by primer extension analysis. Usually, for the analysis, the oligonucleotide is labeled at the 5′-end by γ-[^{32}P]-ATP and polynucleotide kinase. Modified positions close to the 3′-end of the RNA appear as short cDNA sequences (at the bottom of the sequencing gel) while positions modified near the 5′-end of the RNA are reflected by longer cDNA products (in the upper part of the gel). The choice of different oligonucleotide primers allows even very long RNA molecules to be analyzed. The only requirement is that oligonucleotides are available that are complementary to the target RNA at a distance of roughly 200 nucleotides to scan the total sequence of even a very complex RNA molecule. For the exact assignment of the primer extension products on the denaturing polyacrylamide gels a parallel separation of products from a sequence reaction of the non-modified target RNA with all four dideoxy-NTPs (ddNTPs) is commonly performed. A certain challenge for the interpretation of primer extension analyses is caused by sequence-dependent abortions, which may arise from very stable secondary structures or result from the presence of naturally modified nucleotides as they occur in some RNA molecules. A control with non-modified RNA is therefore advisable.

32.8.4 Customary RNases

For the analysis of RNA–protein complexes a set of commercially available purified enzymes is listed in Table 32.1. They can be used universally and cover a broad range of applications for the analysis of RNA structures and RNA–protein complexes (Table 32.1).

RNase T1 RNase T1, isolated from *Aspergillus orizae*, cleaves single-stranded RNA at the 3′ site of guanosine moieties. The phosphate group remains at the 3′ terminal guanosine. The enzyme is relatively insensitive to changes in pH and remains active in the presence of 7 M urea. RNase T1 is a very frequently used enzyme because of its high specificity and limitation to only one base (guanosine). The methylated guanosines m1G and m7G occurring in natural RNA molecules are not recognized.

RNase U2 RNase U2 is isolated from *Ustilago sphaerogena* and cleaves 3′ all four common nucleotide positions in single-stranded RNA with a strong preference for A ($A > G \gg C > U$). Cleaved fragments remain phosphorylated at the 3′ end. RNase U2 has a pH optimum of 4.5 and 7 M urea hardly inhibits the cleavage. For the analysis of most complexes the conditions have to be adjusted to neutral pH, which requires correspondingly higher enzyme concentrations

RNase CL3 RNase CL3 is isolated from chicken liver. The enzyme is preferentially used to cleave at non-paired cytosines. Cleavage does also occur behind A and U but requires considerably longer reaction times or higher enzyme concentrations. In all cases the 3′ phosphorylated cleavage fragments are generated. The presence of magnesium ions or spermidine enhances the activity of the enzyme.

RNase T2 RNase T2 is generally derived from *Aspergillus orizae*. This nuclease cleaves at any nucleotide, even at most of the modified ones. Cleavage results in 3′phosphorylated ends. A slight preference for cleavage at adenosine has been observed. The enzyme is active at slightly acidic and neutral pH but is inhibited by metal ions (Cu^{2+}).

Nuclease S1 This nuclease is likewise isolated from *Aspergillus orizae* and hydrolyzes DNA as well as RNA in sequence-independent manner. Hydrolysis leads to 5′-phosphorylated ends. Nuclease S1 is the major enzyme used for characterization of single-strand sequences. Zn^{2+} ions are essential for the reaction. Cleavage is optimal at pH 4.5 and reactions at neutral pH require correspondingly higher enzyme concentrations.

RNase CVE This enzyme, often also designated V1, is isolated from the venom of the cobra snake *Naja naja oxiana* and cleaves selectively double-strand RNA without notable sequence preference, leading to 5′ phosphorylated ends. For recognition as cleavage site a minimum of 4–6 paired nucleotides are required. Occasionally, non-paired sequences may be cleaved if they are in tightly stacked conformation. The enzyme requires Mg^{2+} ions for activity and is strongly inhibited by EDTA.

32.8.5 Chemcal Modification of RNA–Protein Complexes

Chemical reagents can be applied for the structural characterization of RNA–protein complexes in a similar fashion as described for DNA–protein complexes. Due to different secondary structures of RNA and the coherently different accessibilities to the functional groups of the bases additional reagents are in use to account for the different reactivity of RNA molecules. It is also valid here that the smaller sizes of chemical reagents compared to enzymes enable a tighter intrusion into complex structures, contributing to higher resolution of the analysis. The specificity of the chemical reaction does not only yield information about the nucleotides involved in RNA–protein complex formation. Because the reagents attack in each case very distinct functional positions of the nucleic acid chain information about the accessibility of these positions is provided at an atomic level. Among these positions are the N1 and N2 positions of guanine, the N7 and N1 positions of purine bases, the N3 position of pyrimidines as well as the phosphate groups of the nucleic acid backbone. By combination of reagents it is possible to detect differences responsible for the formation of Watson–Crick base pairs or alternative pairings (e.g., Hoogsteen pairs) or differences resulting from deviations in the sugar-phosphate backbone.

It is important to note that chemical modifications primarily indicate *changes in reactivity*. This is not necessarily identical with changes in the *stereochemical accessibility* of the molecule. Another point demanding caution is that many chemical reagents can also modify functional groups of proteins. It is obligatory, therefore, to test whether the complexes remain intact under the reaction conditions. Moreover, for chemical modification reactions buffers should avoided that contain amino groups; this means Tris-buffers are inappropriate, for instance, and should be replaced by HEPES, phosphate, or cacodylate buffers. On the other hand, chemical modifications are much less dependent on the presence of Mg^{2+} ions or EDTA and in contrast to many enzymes do not have a small pH optima. In contrast to RNases chemical modifications do not directly lead to cleavage of the RNA chain. To localize the modified position either an additional cleavage reaction must be performed or the analysis must be completed by a subsequent primer extension reaction if the modification directs a polymerase abort. The reagents indicated below have been proven to be of special value for the analysis of RNA structures and RNA–protein complexes (Table 32.2; Figure 32.20).

Kethoxal (α-Keto-β-ethoxy-butyraldehyde) Kethoxal reacts specifically only with non-paired guanosines by formation of a five-membered ring including the guanine atoms N1

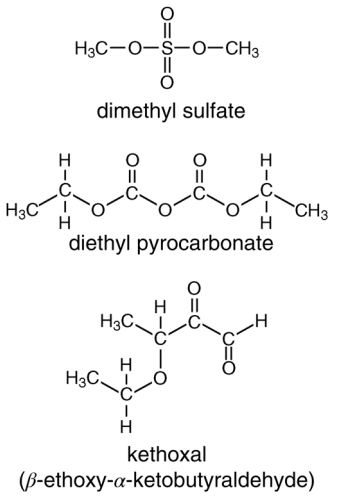


Figure 32.20 Structural formula of RNA-modifying reagents.

Table 32.2 Reagents for the chemical modification of RNA.

Reagent	Specificity	Analysis method[a),b)]	Remarks
Kethoxal	N1-Gua, N2-Gua	PE	Adduct RNase T1-resistant Application *in vivo* possible
DMS	N3-Cyt	Cleavage, PE	Application *in vivo* possible
	N1-Ade	PE	
	N7-Gua	Cleavage, PE after cleavage	
DEPC	N7-Ade	Cleavage, PE	
ENH	Sugar-phosphate backbone	Cleavage, PE after cleavage	Often only applicable as interference method
OH-radicals	Sugar-phosphate backbone	Directly, PE	

a) PE: primer extension.
b) Cleavage: treatment with aniline.

and the amino group of C2. The product is stable at acidic pH and can be stabilized further in the presence of borate. Kethoxal modified guanosines within an RNA chain are no longer attacked by RNase T1, which can be utilized for identification. Because the kethoxal adduct is alkali-labile modified positions within an RNA chain can relatively easy be identified by RNase T1 hydrolyses before and after alkali treatment. Kethoxal, which penetrates through membranes and cell walls, can also be applied for *in vivo* studies with intact cells.

DMS (Dimethyl Sulfate) DMS, equally suited for the modification of DNA, can be employed under special conditions for the selective analysis of accessible positions of guanines (N7-Gua), cytidine (N3-Cyt), and adenine (N1-Ade) within RNA. The reaction of DMS with guanine occurs at the N7 position, introducing a methyl group and a positive charge to the imidazole ring. The slightly destabilized 7,8 double bond can be reduced with Na-borohydride at mild alkaline pH. At the resulting intermediate product (m^7-dihydro guanosine) a strand scission can be performed with aniline. The reaction discloses guanine residues whose structural contribution rests on non-Watson–Crick base pairs. Strong stacking interactions inhibit the modification.

DMS reacts with non-paired cytidines at the N3 position. The modification can be identified by primer extension because the modified cytosine gives rise to a reverse transcription abort. Alternatively, a cleavage reaction can be introduced with end-labeled RNA. This requires a reaction of the modified base with hydrazine prior to the cleavage of the chain by aniline. Note that the cleavage reaction results in unique products with 3′ end-labeled RNA molecules, yielding sharper bands on the subsequent sequencing gel compared to 5′ end-labeled RNAs!

For the analysis of single-stranded adenosines the DMS reaction with N1A, which yields methyl-1-adenine, can be used. The subsequent analysis can only be performed by primer extension because there is no chemical strand scission reaction known for this modification.

DMS reacts with many proteins; hence for the analysis of RNA–protein complexes it is advisable to test the stability of the complexes under the reaction conditions. Because the hydrolysis of DMS in water leads to the formation of sulfuric acid it is important to buffer the decline in pH at high reagent concentrations.

DEPC (Diethyl Pyrocarbonate) The reaction of diethyl pyrocarbonate (DECP) causes a carbethoxylation of the N7 position of purines. Diethyl pyrocarbonate can favorably be used at neutral conditions to test the participation of the N7 position of adenine in tertiary interactions. Watson–Crick type base-paired adenosines in helical regions are not reactive with DECP. The modification is usually identified with end-labeled RNA by an aniline-catalyzed strand scission. At mild acidic pH the N7-Gua position and at slightly alkaline buffer the N3-Ura positions are also carbethoxylated. Like DMS, DECP is known for its reactivity towards proteins, which explains its preferred use for structural analyses of isolated RNA. If RNA–protein complexes are studied their stability should be assured in advance.

CMCT (1-Cyclohexyl-3-(2-morpholinoethyl)carbodiimide metho-*p*-toluolsulfonate) CMCT reacts with uridine (N3) and guanosine (N1) if the respective bases are not paired

according to Watson–Crick. Both modifications can be analyzed by primer extension. For the analysis of RNA–protein complexes CMCT is actually only used in special cases.

ENU (Ethylnitrosourea) This N-nitroso-alkylating reagent attacks the oxygen atoms of the phosphate backbone and thus differs from, for example, DMS, which attacks the ring-nitrogen atoms of the bases. The resulting RNA-phosphotriesters from ethylnitrosourea (ENU) reactions are unstable and can easily be cleaved by mild alkaline conditions. ENU preferentially attacks phosphates oriented in exact helical geometry. The contribution of the phosphates in tertiary interactions or H-bonds to amino acid side chains as well as cation coordination reduces the reactivity. For the analysis of RNA–protein complexes ENU is usually normally only used under damage-selection conditions.

Hydroxyl Radicals The reaction of hydroxyl radicals with DNA–protein complexes has already been described in detail. Hydroxy radicals ($^{\bullet}OH$), generated by the Fenton reaction (32.5) can also successfully be used for the analysis of RNA–protein complexes. The radicals that arise from the reaction of (Fe^{2+})-EDTA complexes and H_2O_2 (32.6) interact with C4′ of the ribose, causing a strand scission. The reaction is independent of the base sequence and only marginally influenced by the secondary structure. Protein contacts or notable changes in the groove geometry, however, cause differences in the modification pattern. The modification pattern is visualized by autoradiography of the sample separated on a sequencing gel, either directly with end-labeled RNA or after a primer extension reaction with a non-labeled RNA. The analysis of hydroxyl radical-modified RNA generally provides information about almost all nucleotides of the molecule and regions of protection or enhanced reactivity can best be identified when the profiles resulting from densitometry of the gel lanes from free- and complexed-RNA are superimposed.

Fe-BABE (Fe 1-(*p*-Bromoacetamidobenzyl)ethylenediaminetetraacetic Acid) The reagent Fe-BABE, conjugated to proteins, has been proven of special value for the analysis of high molecular complex RNA–protein particles. It enables the elucidation of complicated neighborhood relations between RNA molecules and different ligands. Valuable information has been obtained for the topography of ribosomes and RNA polymerases bound to nucleic acids. Because the reagent can be coupled to RNA spatial information on RNA–RNA and RNA–protein interactions can be obtained.

In-Line Probing During the hydrolysis of the 3′-5′-phosphodiester bond of single-strand RNA, contrary to double-strand RNA in the A-form, an arrangement of the 2′-oxyanion position and the neighboring 5′-oxyanion leaving group in the in-line orientation is possible. This orientation significantly facilitates a transesterification reaction and leads to a facilitated spontaneous or metal-catalyzed hydrolysis of non-helical RNA regions. As a result, internucleotide bonds of secondary structured RNA sequences show less frequent spontaneous hydrolysis as RNA sequences, which are unstructured. This fact can successfully be used to distinguish helical from single-stranded RNA sequences. The method, termed in-line probing, is especially effective in deriving structural information on RNA interacting with ligands or metal ions. For the analysis an end-labeled RNA is incubated at room temperature for about 40 h in the presence of Mg^{2+} ions and the products resulting from spontaneous cleavage are separated on denaturing gels and visualized by autoradiography. A limited RNase T1 hydrolysis, showing all guanosine positions, and an alkaline hydrolysis lane, showing statistically all sequence positions of the end-labeled RNA, are generally used for the sequence assignment at the nucleotide level.

SHAPE Analyses A recently developed method for the analysis of RNA secondary and tertiary structures, designated by the acronym SHAPE (selective 2′-hydroxyl acylation analyzed by primer extension), can also be applied to study RNA–protein complexes. The method is based on the difference in reactivity of single-strand versus paired nucleotides of an RNA molecule towards electrophilic reagents like NMIA or 1M7 (1-methyl-7-nitroisatoic anhydride) or benzoyl cyanide (BzCN). The reagents react selectively with the ribose 2′-OH group and local differences in the mobility/flexibility of the nucleotides can be recorded because less flexible nucleotides, as they exist in single-strand RNA, are found more frequently in a conformation that is more reactive

towards electrophilic reagents. As a result, nucleotides of single-strand regions or nucleotides not fixed by protein contact show stronger modification by the SHAPE reagents compared to nucleotides involved in stable base pairing or fixed by tertiary interactions. Because the SHAPE-modification causes an abortion of the cDNA synthesis the modified positions can readily be identified by primer extension reaction. The method provides information on the nucleotide level about the structure or dynamics of RNA molecules or RNA–protein complexes. It can furthermore be performed by high-throughput technology (hSHAPE) in which different fluorescence labeled primers are used and the resulting cDNA products are separated by capillary electrophoresis.

32.8.6 Chemical Crosslinking

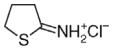
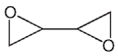
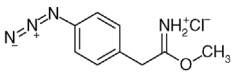
Chemical Crosslinking, Section 6.3

For the identification of contact sites between binding partners chemical crosslinking is still a method of choice. As an advantage, crosslinking can convert weak, non-covalent interactions between proteins and nucleic acids into covalent and chemically stable conjugates. For that reason even complexes of low stability and short lifetime can be characterized by crosslinking. Note, however, that with crosslinking only positive results can be interpreted because the absence of a crosslink may be due to chemical or steric reasons and must not rule out a direct contact! For crosslink analyses different bifunctional reagents are available that possess protein-specific as well as nucleic acid-specific reactive groups (hetero-bifunctional reagents). Each crosslinking reagent has a characteristic distance of its functional groups (Table 32.3). This distance limits the precision of the analysis. Because only information of closely neighboring functional groups is desired from crosslinking experiments, preferably reagents with short distances (about 1 nm) are considered. Nucleic acids can also be crosslinked by direct irradiation with energy-rich UV light between 250 and 280 nm. This type of crosslinking occurs at virtually zero-length.

Crosslink reactions often suffer from low yields, which sometimes enforces more drastic reaction conditions. Crosslinks from such studies are not always restricted to the protein–nucleic acid contact (intermolecular) and numerous links within the nucleic acid (intramolecular), degradation caused by photo-cleavage or oxidations, have to be accepted.

This can in many cases be avoided by the use of crosslinking reagents with photoreactive groups that are activated by mild UV-radiation (>300 nm). 2-Iminothiolane is such a reagent, for instance. Photoreactive reagents can also be used for stepwise crosslinks with a first chemical reaction to functional groups of a binding protein followed by a second reaction with the nucleic acid after formation of the protein–nucleic acid complex initiated by mild UV radiation. There are several different strategies for subsequent localization of the contact points at either the protein or the nucleic acid level. The original very laborious conventional peptide fingerprint methods used to identify the amino acids involved have been replaced by fast and sensitive mass spectrometric techniques (MS) allowing a high sample throughput and work with protein amounts in the picomole range. Basically, the crosslinked complex is hydrolyzed by nucleases and proteases. The protein can, for instance, be digested with endo-

Table 32.3 Some crosslinking reagents and their properties.

Reagents	Structure	Reaction	Protein	RNA	Distance (nm)
2-Iminothiolane		UV (350 nm)	$-NH_2$	Uridine	0.5
Diepoxybutane (DEP)		37 °C, pH 7	$-NH_2$(Lys) –SH (Cys)	N7-G	0.46
Methyl-*p*-azidophenyl-acetimidate (APAI)		1. Tris buffer 2. 300–460 nm	$-NH_2$ (Lys)	Non-specific	ca. 1
Formaldehyde	$H_2C{=}O$	37 °C, pH 8	Lys, Arg, His, Trp	C, G, A, T; Amino- and imino groups	0.2

peptidase Lys-C, while nucleic acids (in case of RNA) are hydrolyzed by RNase T1. The resulting peptide-oligonucleotide adducts are then purified by column chromatography or gel electrophoresis. The purified hetero-conjugates are subsequently sequenced by matrix-assisted laser desorption/ionization time of flight (MALDI-TOF) spectrometry, which allows determination of the sequences of both the peptide and the nucleic acid and identification of the point of the crosslink. With chemical crosslinking reagents, and not direct UV-cross-linking, the exact mechanism of the crosslinking reaction must be known to derive the exact mass of the hetero-conjugate consisting of the peptide moiety, crosslinking reagent, and oligonucleotide.

32.8.7 Incorporation of Photoreactive Nucleotides

In place of using bifunctional reagents or non-directional UV-irradiation, crosslinking can also be achieved by the directed incorporation of photoreactive nucleotides. A set of different nucleotide analogs is available, which can either be integrated into the nucleic acid chain by enzymatic incorporation or by chemical synthesis. They may alternatively be added to each of the 3′ or 5′ ends. For the analysis of protein–nucleic acid complexes the incorporation of *^{4}S-UTP* has been of value. This nucleotide analog is accepted by T7 RNA polymerase and can be positioned by *in vitro* transcription reactions instead of UTP, either statistically or following suitable transcription strategies, at defined positions within an RNA chain. ^{4}S-uracil has the advantage of behaving very similarly to uracil and no notable changes in the structure of the nucleic acid must be expected. The excitation wavelength for crosslinking (suitable for RNA with and without bound protein) is 300 nm, which does not create a risk for the integrity of RNA–protein complexes. The crosslink positions within RNA can generally be identified by comparative primer extension reactions with radiated and non-irradiated RNAs.

Photoreactive nucleotide derivatives likewise exist for the directed incorporation into DNA, enabling crosslinking analyses of DNA–protein complexes as well. The strategies for incorporation differ here, of course. Derivatized oligonucleotides, serving as primers for a PCR reaction, which can place the reactive nucleotides at any desired position within a DNA fragment, are the matter of choice. Alternatively, crosslinking nucleotides can be placed at 5′ overlapping ends of DNA fragments with appropriate DNA polymerase fill-in reactions and photoreactive deoxynucleotide triphosphates as substrates. The fragment can be flanked by a second DNA fragment by a subsequent ligation reaction.

For the choice of reaction conditions it must be considered that some reagents are light-sensitive and all steps of the procedure have to be performed under darkened laboratory conditions. Some photoreactions are disturbed by the presence of reducing agents. Photoreactive azides, for example, can be reduced to the non-reactive amines. Among the interfering substances are also the thiol reagents 2-mercaptoethanol or DTT. These substances, often added for the protection of SH-groups in proteins, should therefore not exceed a concentration of 50–100 μM.

32.8.8 Genome-Wide Identification of Transcription Start Sites (TSS)

The quantitative determination of the total RNA from an organism by microarray technology is a very useful method for the characterization of changes in expression on the RNA level caused by alterations of the external conditions or as a consequence of mutations (transcriptome analyses). Today, the identification of transcripts relies increasingly on high-throughput sequencing techniques (RNA-Seq) instead of hybridization-based methods. The RNA-Seq procedure is based on the translation of the total RNA preparation into a library of shorter cDNA fragments. To these fragments sequencing adaptors are ligated and the sequence is established by current high-throughput techniques (e.g., 454 pyrosequencing). The initially obtained short sequences (sequence reads) are aligned with the reference genome, which often enables the establishment of a transcription profile of all genes from the organism at nucleotide resolution.

Transcriptome Analysis, Section 37.1

High- throughput Sequencing Techniques, Section 30.2.1

Conventional transcriptome analyses generally record the steady state concentration of all present RNAs as long as they are represented on the microarray chips. It is not possible to distinguish between primary transcripts and processed RNAs. There are, however, suitable methods to

determine, on a genome-wide basis, direct changes of the primary transcripts allowing us to map and to annotate accordingly a primary transcriptome. If the exact transcription start sites are known microarrays with overlapping probes (tiling arrays) can be used to quantify the RNA level in direct proximity to the promoter. In eukaryotic systems primary transcripts can also be identified by their 5′ Cap structure and suitable procedures for the specific enrichment and identification of Cap structures exist (Cap analysis of gene expression, CAGE). For prokaryotic transcripts, which do not have a Cap, advantage can be taken of the fact that prokaryotic primary transcripts generally have a 5′ triphosphate, whereas the 5′ end of processed RNA is characterized by a 5′ monophosphate. For a TTS determination in prokaryotic systems, tiling arrays and RNA-Seq procedures are combined with a technique that allows distinguishing between RNAs with a 5′ triphosphate or a 5′ monophosphate moiety. For this procedure two different cDNA libraries are established. The first library is prepared from non-treated total RNA while the RNA for the second library is treated with terminator exonuclease (TEX), which digests RNA with 5′ monophosphates from their 5′ ends but not RNA with a 5′ triphosphate or a free 5′ OH group (or RNA with a 5′ Cap). Enriched primary transcripts, obtained in this way, are then identified by RNA sequencing techniques (454 Pyrosequencing, RNA-seq). The method is designated as differential RNA-seq (dRNA-seq) and allows genome-wide localization of RNA polymerase start sites (TTS or promoter regions).

32.9 Genetic Methods

32.9.1 Tri-hybrid Method

Two-Hybrid System, Section 16.1

Like the two-hybrid system, serving for the identification of interacting proteins within the cell, an analog procedure, the tri-hybrid system, serves to identify RNA-binding proteins. The principle is very similar to the two-hybrid system. This is based on the activation of reporter genes within the cell, for instance by the transcriptional activator GAL4. This factor is expressed in the cell by special expression vectors as two separate fusion proteins. One of these proteins contains the activation domain, the other the DNA-binding domain. When both proteins interact the transcription factor activity is regenerated and the reporter gene can be transcribed. In this way, the two-hybrid system serves to demonstrate protein–protein interactions. The tri-hybrid system goes one step further. Now two fusion proteins are expressed, which in this case consist for instance of the GAL4 DNA-binding domain and the protein RevM10 (a mutated HIV-Rev protein) as well as the GAL4 activation domain and a putative RNA binding protein. In addition, an RNA hybrid, consisting of a binding sequence for the RevM10 protein (RRE) and the RNA to be examined, is transcribed in the cell by RNA polymerase II. The activation of the reporter gene depends exclusively on the binding of the RNA to be studied and the putative RNA-binding protein by which the functional elements for transcription activation are brought together (Figure 32.21). Alternatively, the tri-hybrid method can make use of fusion proteins from bacterial transcription factors when the reporter

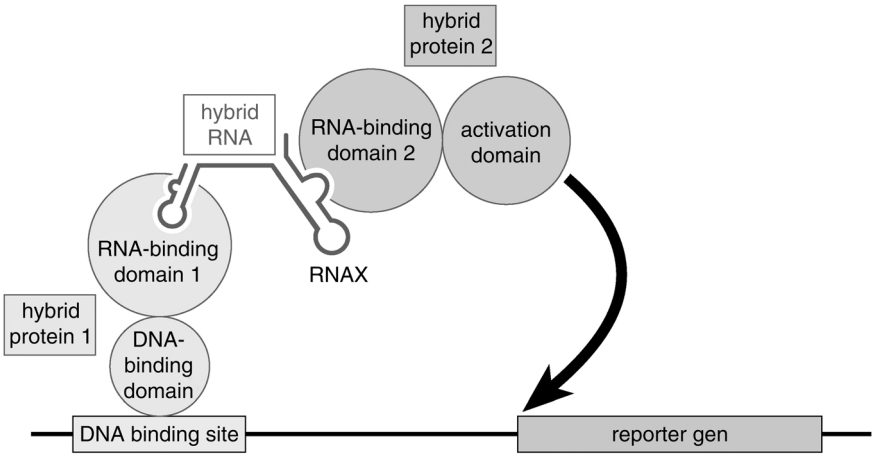


Figure 32.21 Scheme of a tri-hybrid system for the *in vivo* characterization of RNA–protein recognition. Source: according to Bernstein, S. *et al.* (2002) *Methods*, **26**, 123–141. With permission, Copyright © 2002 Elsevier Science (USA).

gene contains the respective regulatory sequences in the vicinity of the promoter. For instance, for the LexA operator, a fusion protein of the LexA repressor and the MS2 phage coat protein as well as a fusion between the iron-regulatory protein 1 can be expressed. As hybrid RNA, a fusion with binding sites for the MS2 coat protein and the iron-responsive element IRE can be used. In addition, systems have been described that take advantage of the interaction between HIV transactivator protein (Tat) and an RNA with the HIV trans-activator response element (TAR).

Generally, tri-hybrid systems are especially appropriate to identify specific RNA-binding proteins or target RNAs *in vivo* from a library of cDNAs or to analyze structural peculiarities of an already identified RNA–protein interaction. In the latter case mutant RNA molecules are expressed. There also exists a one-hybrid screen for the identification of specific DNA sequences. In this case the DNA sequences are cloned upstream of a yeast reporter gene (target element).

The proteins to be analyzed are cloned as fusion proteins with an activation domain for the reporter gene. A convenient combination is for instance the GAL4 activation domain with its recognition sequence and β-galactosidase or HIS3 as reporter gene. For screening experiments a yeast strain is constructed that has this reporter gene with the upstream positioned β-galactosidase integrated into its genome (reporter strain). This strain is then transformed with a cDNA library of candidate genes containing the protein sequences to be analyzed as fusion into the activation domain (GAL4 system). Positive candidates can be identified by a color reaction on corresponding agar plates.

32.9.2 Aptamers and the Selex Procedure

Selection of Aptamers (SELEX), Section 38.4.1

The Selex (systematic evolution of ligands by exponential enrichment) method combines the possibilities of biochemical *in vitro* synthesis with the potential of genetic selection. It is a matter of *in vitro* selection of RNA (or DNA) molecules that are roughly 20 to 60 nucleotides long with specific properties, such as binding of small ligands, protein binding, or catalytic activity, from an enormous pool of random sequences. The Selex procedure has been used successfully in the past for the characterization of protein binding sites, binding of various different ligands to nucleic acids, as well as for the construction of ribozymes. The selected nucleic acids exhibit the properties of highly specific receptors and are designated aptamers. The term aptamer (Latin: *aptus*, to fit) depicts that out of the huge number of non-functional molecules ($>10^{15}$) those with a suitable function were enriched. At first a random library of DNA or RNA molecules containing as much as possible randomized sequences in their middle are generated. The ends of the molecules in contrast contain known sequences, which serve as primer targets for PCR amplification. The evolutionary Selex method is based on the separation of those molecules from the total pool that exhibit binding properties (or catalytic activity). This is for instance achieved by the use of an affinity column. Those molecules with the best retention are used for further rounds of selection. In the case of RNA they are subjected to reverse transcription, converted into double-strands, and amplified by PCR. Usually, the ends of the molecules contain a sequence for the phage T7 RNA polymerase promoter, which facilitates the synthesis of correspondingly large amounts of the enriched single-strand RNA. From this pool of RNAs the fraction of functional molecules is separated again by a new binding reaction (or test for catalytic activity) and the cycle is repeated until no better binding properties can be observed (about 10–20 cycles). The stringency for binding may be enhanced between single cycles or randomization may be enhanced again by deliberately inaccurate cDNA synthesis enhancing the evolution rate. Usually, the number of sequence variants is limited in the start pool. For RNA with a length of 220 nucleotides there are theoretically $4^{220} = 10^{132}$ sequences possible. Such an enormous amount of RNA cannot be realized. Generally, one starts with 10^{15} sequences. With the previous randomization the active pool can be evolved. With the Selex procedure numerous nucleotide sequences with amazing properties have been generated. These include molecules with a specific binding capacity for small biomolecules, like amino acids, sugars, dyes, or antibiotics. In addition, aptamer molecules with different catalytic properties, not yet observed in nature, have been generated. Of special interest for the analysis of protein–nucleic acids is the fact that RNA-binding motifs for specific proteins, for instance the HIV-Rev protein, could be evolved. Sequence comparison and the measured binding constants allow conclusions to be made about the specific structural elements involved in protein binding.

32.9.3 Directed Mutations within Binding Domains

When an RNA–protein binding partner (or likewise a DNA–protein binding partner) has been identified and the binding sequence within the nucleic acid has been localized subsequent directed or statistical *in vitro* mutagenesis procedures combined with binding experiments can yield a very precise description of the molecular basis of the protein–nucleic acid recognition. Within recent years the methods for targeted or comprehensive mutagenesis have been refined and simplified. The use of oligonucleotides has become a standard application for the molecular biologist and numerous commercially available kits facilitate fast, PCR-based mutagenesis techniques.

For a systematic change of preferably all nucleotides or amino acids potentially involved in the interaction a mutagenesis method designated linker-scanning or alanine-scanning has been developed. In the linker-scanning mutagenesis a fixed sequence (linker) is used for the systematic replacement of nucleotide sequences over a defined sequence region. In the method originally developed two nested series of 5′ and 3′ deletions starting from a restriction enzyme recognition site of the target DNA are created at first. At each end of the deletion construct the same short oligonucleotide (linker) is ligated. Subsequently, both sets of 3′ and 5′ truncated deletion constructs leading to the original length are ligated, whereby the linker sequence is located at a different position in each newly formed fragment. As a result, the position of a distinct sequence is permutated over a large range. In this way position effects of binding sites can be varied systematically. The isolated constructs can then be used for binding studies with any assay and the quality of complex formation tested with each separate construct.

In the alanine-scanning method codons within the mRNA of a binding protein of interest are systematically replaced with sequences encoding the neutral amino acid alanine (GCU, GCA, GCC, GCG). This is achieved by standard mutagenesis procedures. The gene sequence must of course be known and accessible for mutagenesis. The mutated sequences have to be cloned in an expression vector and potential effects on binding can be analyzed *in vivo* after transformation of a suitable strain by a screening procedure (tri-hybrid assay). Frequently, the altered proteins should be studied directly. This is of course a laborious step. It is advisable to choose an expression vector that adds a His-tag to the mutant protein gene. This enables a one-step enrichment of the expressed protein variants from the cell extract.

If the point is to learn the molecular details of a distinct binding mechanism generally only target-directed mutations are created. For instance, charged or hydrophobic amino acids would be replaced by neutral ones. In the same manner, binding regions of nucleic acid components can also be mutated by single-nucleotide replacements, of course. In this case, the already existing information on the nature of the binding site is essential. For studies like this, Selex experiments can provide excellent information, for instance. In any case, the *in vitro* mutagenesis of binding sites in combination with binding experiments represent simple but very powerful methods with a great potential to elucidate the molecular mechanisms of protein–nucleic acid interactions.

Further Reading

Bernstein, D., Buter, N., Stumpf, C., and Wickens, M. (2002) Analyzing mRNA–protein complexes using a yeast three-hybrid system: methods and applications. *Methods Enzymol.*, **2**, 123–141.

Bustamante, C. and Rivetti, C. (1996) Visualizing protein–nucleic acid interactions on a large scale with the scanning force microscope. *Annu. Rev. Biophys. Biomol. Struct.*, **25**, 395–429.

Dame, R.T., Wyman, C., and Goosen, N. (2001) Structural basis for preferential binding of H-NS to curved DNA. *Biochimie*, **83** (2), 231–234.

Draper, D.E. (1995) Protein-RNA recognition. *Annu. Rev. Biochem.*, **64**, 593–620.

Famulok, M. and Verma, S. (2002) *In vivo*-applied functional RNAs as tools in proteomics and genomic research. *Trends Biotechnol.*, **20**, 462–466.

Favre, A., Saintome, C., Fourrey, J.L., Clivio, P., and Laugaa, P. (1998) Thionucleobases as intrinsic photoaffinity probes of nucleic acid structure and nucleic acid-protein interactions. *J. Photochem. Photobiol. B*, **42** (2), 109–124.

Fried, M.G. and Daugherty, M.A. (1998) Electrophoretic analysis of multiple protein–DNA interactions. *Electrophoresis*, **19** (8–9), 1247–1253.

Hermann, T. and Westhof, E. (1999) Non-Watson–Crick base pairs in RNA–protein recognition. *Chem. Biol.*, **6** (12), 335–343.

Hill, J.J. and Royer, C.A. (1997) Fluorescence approaches to study of protein–nucleic acid complexation. *Methods Enzymol.*, **278**, 390–416.
Huber, P. (1993) Chemical nucleases: their use in studying RNA structure and RNA–protein interactions. *FASEB J.*, **7**, 1367–1374.
Jensen, O.N., Kulkarni, S., Aldrich, J.V., and Barofsky, D.F. (1996) Characterization of peptide-oligonucleotide heteroconjugates by mass spectrometry. *Nucleic Acids Res.*, **24**, 3866–3872.
Lane, D., Prentki, P., and Chandler, M. (1992) Use of gel retardation to analyze protein–nucleic acid interactions. *Microbiol. Rev.*, **56** (4), 509–528.
Moine, H., Cachia, C., Westhof, E., Ehresmann, B., and Ehresmann, C. (1997) The RNA binding site of S8 ribosomal protein of Escherichia coli: Selex and hydroxyl radical probing studies. *RNA*, **3** (3), 255–268.
Mukherjee, S. and Sousa, R. (2003) Use of site-specifically tethered chemical nucleases to study macromolecular reactions. *Biol. Procedures Online*, **5**, 78–89.
Pavski V. and Le, X.C. (2003) Ultrasensitive protein–DNA binding assays. *Curr. Opin. Biotechnol.*, **14** (1), 65–73.
Putz, U., Skehel, E., and Kuhl, D. (1996) A tri-hybrid system for the analysis and detection of RNA–protein interactions. *Nucleic Acids Res.*, **24**, 4838–4840.
Record, M.T., Jr, Zhang, W., and Anderson, C.F. (1998) Analysis of effects of salts and uncharged solutes on protein and nucleic acid equilibria and processes: a practical guide to recognizing and interpreting polyelectrolyte effects, Hofmeister effects, and osmotic effects of salts. *Adv. Protein Chem.*, **51**, 281–353.
Schulz, A., Mücke, N., Langowski, J., and Rippe, K. (1998) Scanning force microscopy of *Escherichia coli* RNA polymerase sigma54 holoenzyme complexes with buffer and in air. *J. Mol. Biol.*, **283**, 821–836.
Senear, D.F. and Brenowitz, M. (1991) Determination of binding constants for cooperative site-specific protein-DNA interactions using the gel mobility-shift assay. *J. Biol. Chem.*, **266** (21), 13661–13671.
Soukup, G.A. and Breaker, R.R. (1999) Relationship between internucleotide linkage geometry and the stability of RNA. *RNA*, **5**, 1308–1325.
Steen, H. and Jensen, O.N. (2002) Analysis of protein–nucleic acid interactions by photochemical crosslinking and mass spectrometry. *Mass Spectrom. Rev.*, **21** (3), 163–182.
Tan, W., Wang, K., and Drake, T.J. (2004) Molecular beacons. *Curr. Opin. Chem. Biol.*, **8** (5), 547–553.
Thiede, B., Urlaub, H., Neubauer, H., Grelle, G., and Witmann-Liebold, B. (1998) Precise determination of RNA–protein contact sites in the 50S ribosomal subunit of *Escherichia coli*. *Biochem. J.*, **334**, 39–42.
Wower, I., Wower, J., Meinke, M., and Brimacombe, R. (1981) The use of 2-iminothiolane as an RNA–protein crosslinking agent in *Escherichia coli* ribosomes, and the isolation on 23S RNA of sites crosslinked to proteins L4, L6, L23, L27, and L29. *Nucleic Acids Res.*, **9**, 4285–4302.

Part V

Functional and Systems Analytics

Sequence Data Analysis

33

Boris Steipe

University of Toronto, Department of Biochemistry, 1 King's College Circle, Toronto, Ontario M5S 1A8, Canada

33.1 Sequence Analysis and Bioinformatics

Over less than the last half century, the life sciences have undergone their most radical transformation yet. We have set out from the detailed study of individual molecules, to arrive in the genomic era, and we are now setting our sights on a post-genomic era in which the heritable information that makes up the identity of an organism can simply be assumed as. Several paradigm shifts have accompanied this development: in the past, we focused on developing algorithms that interpret sequences, more recently the challenge has been to integrate information across a large number of data sources that are freely available on the Internet. The present discussion in the field is focused on the challenges of "Big Data." In the past, we have programmed institutional mainframe computers in compiled languages such as Fortran and C, later the focus shifted to Perl, PHP, and other interpreted languages to assemble data from the Internet on our desktop computers, and currently the emphasis is on large libraries of functions as in the statistical workbench **R**, and storage and computation is shifting into the *cloud*. This chapter will introduce all three paradigms: sequence information, integration, and genome-scale analysis, providing pointers to practical approaches wherever possible.

DNA Sequencing, Chapter 30

With industrial-scale sequencers now available for less than the cost of a dump truck, sequencing has gone mainstream and the number of sequenced genomes is still growing exponentially: the US National Center for Biotechnology Information (NCBI) lists more than 4900 sequenced eukaryotic genomes and more than 125 000 sequenced prokaryotes as of December 2017.

The management of such data volumes is not trivial. We are fortunate, however, that a massive reduction of data storage costs in the last decade has approximately kept pace with the deluge of data. The trend to delocalize storage into the *cloud* – globally distributed commodity data centers that store data efficiently and securely for a small fee – is growing and is reducing the cost of data management further. And the data is public, and freely available.

Two major institutions are global hubs of databases and online services: the NCBI (http://www.ncbi.nlm.ncbi.gov) and the European Bioinformatics Institute (EBI, http://ebi.org) in England. These exchange data daily through a data sharing agreement that also includes the DNA Data Bank of Japan (DDBJ), manage a large set of related and extensively cross-referenced databases, and run powerful data-analysis centers to freely support public queries. Beyond that, literally thousands of smaller, online data resources and services are available. Certainly, solving most practical sequence analysis problems today requires nothing more than an Internet connection.

Sequence analysis is located roughly between two poles. On one hand we have bioinformatics in its narrow sense: the technologies behind the management of large datasets,

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

search, retrieval, consistency, and cross-referencing. On the other hand, we have computational biology: the abstraction of biomolecules and their study with computational methods. This is reflected in the topic of this chapter, biological sequence analysis. Biological macromolecules are heterocopolymers of nucleotide or amino acid units. This makes it simple to define an abstraction that is ideally suited for computational analysis. Each unit is described with a letter of the alphabet and thereby a complex biomolecule is mapped to a simple string, which can be efficiently stored, manipulated, and retrieved. But we must not forget that such abstractions lose information – for example, sequences do not represent post-translational modifications, or structural conformers, and thus the representation of biology is incomplete. Sequences are models of molecules, and we need to be aware of the models' limitations as well as the necessity to relate our computational results back to the molecules they represent.

33.2 Sequence: An Abstraction for Biomolecules

It is impossible to develop an intuition for biological sequences if one is not familiar with the one-letter code that is used to map amino acids to characters (Table 33.1). One of the most important procedures of sequence analysis is to evaluate sequence changes, and requires relating sequence changes to a change in molecular properties. Without the mapping sequence back to molecules through the one-letter code, we may be doing informatics but it is not "bio"-informatics.

Nucleotide sequence information basically resides mostly in the identity or non-identity of a nucleotide, and perhaps the complementarity of C/G and A/T pairs. Treating nucleotide sequences merely as abstract strings is a reasonable approximation.

Table 33.1 The 20 proteinogenic amino acids and their one-letter code (according to IUPAC-IUB). Compare this with Appendix 1 regarding the amino acids' biophysical properties and Appendix 3 for the chemical structure.

Code	Amino acid	Mnemonic
A	Alanine	A-lanine
C	Cysteine	C-ysteine
D	Aspartic acid	aspar-D-ic acid
E	Glutamic acid	glut-E-mic acid
F	Phenylalanine	F-enylalanine
G	Glycine	G-lycine
H	Histidine	H-istidine
I	Isoleucine	I-soleucine
K	Lysine	Y-turned-sideways resembles K
L	Leucine	L-eucine
M	Methionine	M-ethionine
N	Asparagine	asparagi-N-e
P	Proline	P-roline
Q	Glutamine	"Q"-tamine
R	Arginine	"R"-ginine
S	Serine	S-erine
T	Threonine	T-hreonine
V	Valine	V-aline
W	Tryptophan	with a lisp: t-W-yptophan
Y	Tyrosine	t-Y-rosine

33.3 Internet Databases and Services

Several factors have led to a virtual explosion of biological databases and services on the Web. These include:

- a large volume of sequence and related data;
- inexpensive computer storage, processing, and Internet connection;
- the availability of free, powerful, well-documented database back-ends for all common computer platforms;
- the availability of free, powerful, well-documented Web servers, such as Apache (http://apache.org/);
- the ease of use of scripting and programming languages such as **R** (http://r-project.org), Perl (http://perl.org), Python (http://python.org/), PHP (http://php.net), and JavaScript (automatically included in all modern Web-browsers) that have excellent support to tie databases to dynamically created Web pages;
- the availability of important algorithms as free and open software, supported through a large user-base that maintains and develops the large libraries of functions, most notably Biopython (http://biopython.org, although with less comprehensive function libraries for bioinformatics than for many other fields), and **R** (http://r-project.org) with the associated Bioconductor project (http://bioconductor.org), which has currently the most active community of open-source bioinformatics developers and by far the most comprehensive code base.

These factors have led to unprecedented opportunities for small- and medium-sized research groups to create and publish their own data resources, to make them available on the Internet, and to connect them to other databases. While this is a marvelous development in principle, it does raise the question of maintaining an overview over which resources exist, how well they are updated and maintained, and what is currently considered to be the state-of-the-art. Beyond the NCBI and the EBI and the larger, community maintained model-organism databases, evaluating a given online resource is a non-trivial task for the non-expert. Which alternatives exist? How are the resources generated exactly? How many false-positive or false-negative results are to be expected? How often is the resource updated? Is the long-term, stable existence of the resource ensured? What is the level of user support and documentation? It is obvious that reproducible research requires confidence in these issues, but it is certainly true that not all online resources can satisfy these requirements and that nearly all have room for improvement.

There are several resources that help maintain a rough overview: *Nucleic Acids Research* (*NAR*, https://academic.oup.com/nar) publishes an annual issue of peer-reviewed database resources and of Web services. These are collected, organized by keyword, and searchable online in the curated Bioinformatics Links Directory of Bioinformatics.ca (https://bioinformatics.ca/links_directory/). Besides *NAR*, bioinformatics articles are frequently published in *Bioinformatics* (https://academic.oup.com/bioinformatics), *BioMed Central Bioinformatics* (https://bmcbioinformatics.biomedcentral.com), and more than 170 other journals that can be found as the result of a keyword search for "bioinformatics" at the US National Library of Medicine (https://www.ncbi.nlm.nih.gov/nlmcatalog). More information about current developments can be found in conference abstracts, for example, the annual Intelligent Systems in Molecular Biology Conference (ISMB, https://www.iscb.org/) of the International Society for Computational Biology. Finally, dedicated online forums exist that will help find answers to practical questions about current best-practice. The most active one specifically for bioinformatics is Biostars (https://biostars.org), but many related questions have been discussed (and answered) on one of the invaluable StackExchange forums (https://stackexchange.com/); Quora is a new addition, often with very high-quality answers (https://www.quora.com/topic/Bioinformatics).

To solve a specific problem with the best currently available tools, most likely the best advice is to follow your peers. Find a relevant recent publication – and in this field *recent* means no older than two or three years – in a well-reviewed journal, and study the methods section carefully. Obviously, the paper's authors are usually pleased to pass on advice.

33.3.1 Sequence Retrieval from Public Databases

The first step of sequence analysis is obviously to obtain the sequence. Sequences may come from in-house projects, or downloaded from public databases. Sequences can be retrieved via a feature search (e.g., searching for human hemoglobin), or more specifically via accession numbers. Searches in general are very well integrated across the various NCBI databases, using the Gquery interface to the *Entrez* database system (Figure 33.1). This is an interconnected system of cross-referenced databases that can be accessed through a unified interface. For example, a search for the Mbp1 transcription factor in baker's yeast yields links into literature, nucleotide, protein, and structure databases, cross references to the yeast genome, sequences in related organisms, expression profiles, and more (Table 33.2).

All database entries – sequences or otherwise – are identified by one or more unique accession numbers. In principle, such accession numbers can be used to cross-reference data between databases. However, one needs to consider whether (biological) objects that are identified by the same accession number are in fact identical; this depends on the exact database semantics and data management policies. In practice, databases usually maintain their own system of accession numbers, and mapping – translating between them – is often necessary. Web services for accession number mapping can be found at the UniProt ID mapping service

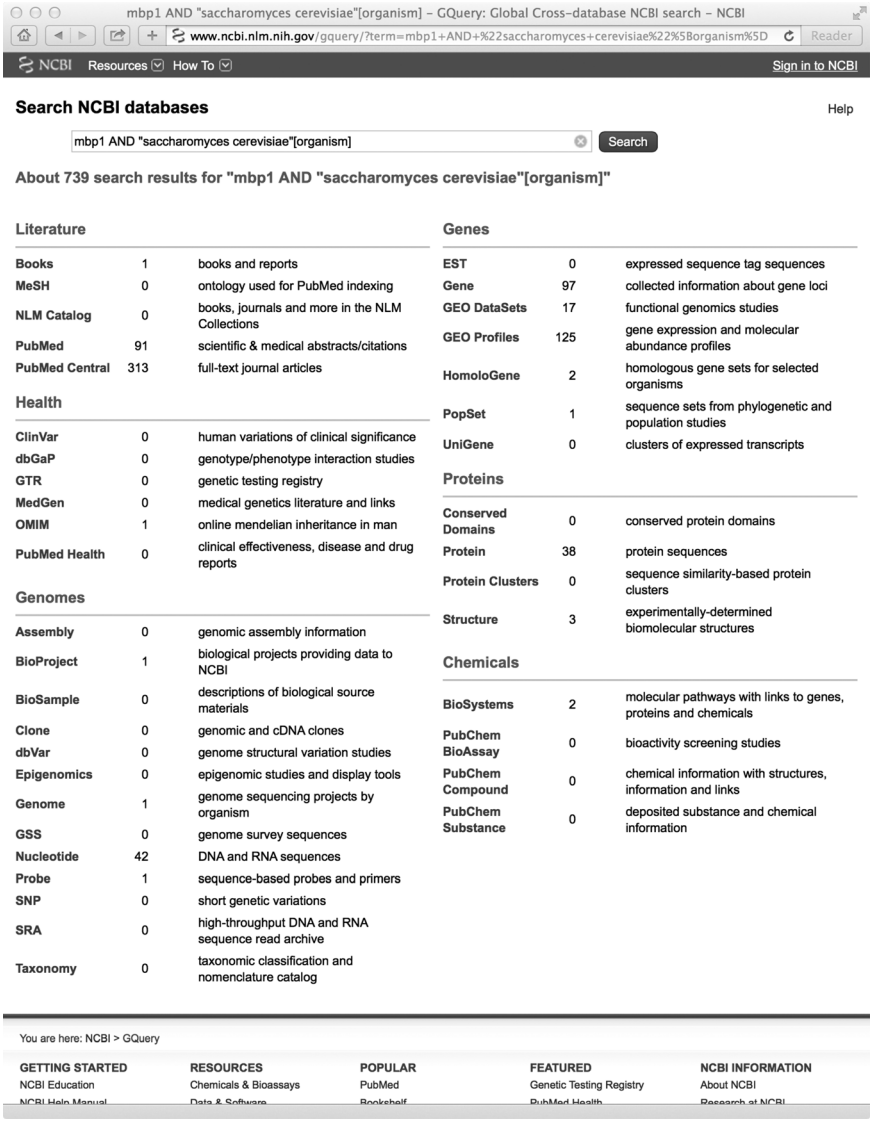


Figure 33.1 Result of a keyword search for the Mbp1 protein (by text-word) in *Saccharomyces cerevisiae* (by organism) at the NCBI. This is a quick and comprehensive way to obtain cross-referenced results. In this example, we see links to journal articles, nucleotide and protein sequences, gene expression profiles, protein structures, and more.

Table 33.2 Selected accession numbers and their semantics for the yeast Mbp1 transcription factors.

	Accession number	Semantic
Name	MBP1	Standard name
Gene	YDL056W	Systematic gene name
	851503	NCBI Entrez Gene ID
	NM_001180115	RefSeq nucleotide (m-RNA) ID
	X74158.1	Nucleotide accession number of the gene (NCBI GenBank and European Nucleotide Archive), version 1
	GI:296143308	GeneInfo – NCBI internal accession number for the gene
Protein	MBP1_YEAST	Swiss-Prot name
	NP_010227	RefSeq protein ID
	P39678	UniProt (protein) identifier
	GI:6320147	GeneInfo – NCBI internal accession number for the protein
Annotation cross-references	S000002214	*Saccharomyces* Genome Database (SGD) ID
	sce:YDL056W	KEGG pathway database ID
	1BM8	PDB (protein structure database) ID for the DNA binding domain
	16823090	iREF protein interaction database ID

(http://www.uniprot.org/mapping/). Uploading single or multiple accession numbers translates them into their corresponding identifiers at the EBI, the NCBI, organism-specific databases, and many others. Probably the most exhaustive translations are provided by bioDBnet (https://biodbnet-abcc.ncifcrf.gov/db/db2db.php), which covers more than 200 databases with more than 700 mappings between them.

But sequences do not only need to be mapped between individual providers' resources, the major providers may have the same information in several different databases. One example is the NCBI's *refseq* project, a curated database designed to solve the problem that primary databases may contain large numbers of redundant, identical sequences from distinct submissions (https://www.ncbi.nlm.nih.gov/refseq), by storing only non-redundant, well-annotated sequences. A similar purpose is fulfilled by the Swiss-Prot subset of the UniProtKB databases – a collection of curated, reviewed, and manually annotated sequences, probably the gold-standard for biological sequence quality (https://www.uniprot.org). Moreover, UniProt offers the option to retrieve only sequences that exceed a selectable threshold of sequence differences between each other, to be able to remove near-identical sequences from a result.

33.3.2 Data Contents and File Format

Modern data technologies often work with well-defined data-grammars, such as XML or JSON. However, bioinformatics is largely still the domain of so-called flat-file formats. Despite difficulties in writing consistent file-parsers, or even accessing precise specifications in the first place, this field sees the advantages of nimble, human-readable formats that are conceptually easy to handle and which can be read and produced by a very large number of legacy applications (Figure 33.2). Such flat-files usually contain *records* of information, prepended with an identifier. Database entries that are displayed on the Web usually contain linked cross-references and graphical elements – the raw information can be downloaded as text.

Since most analyses require only the sequence itself, however, the FASTA format is generally adopted as the *de facto* standard for data interchange (Figure 33.3). It consists of

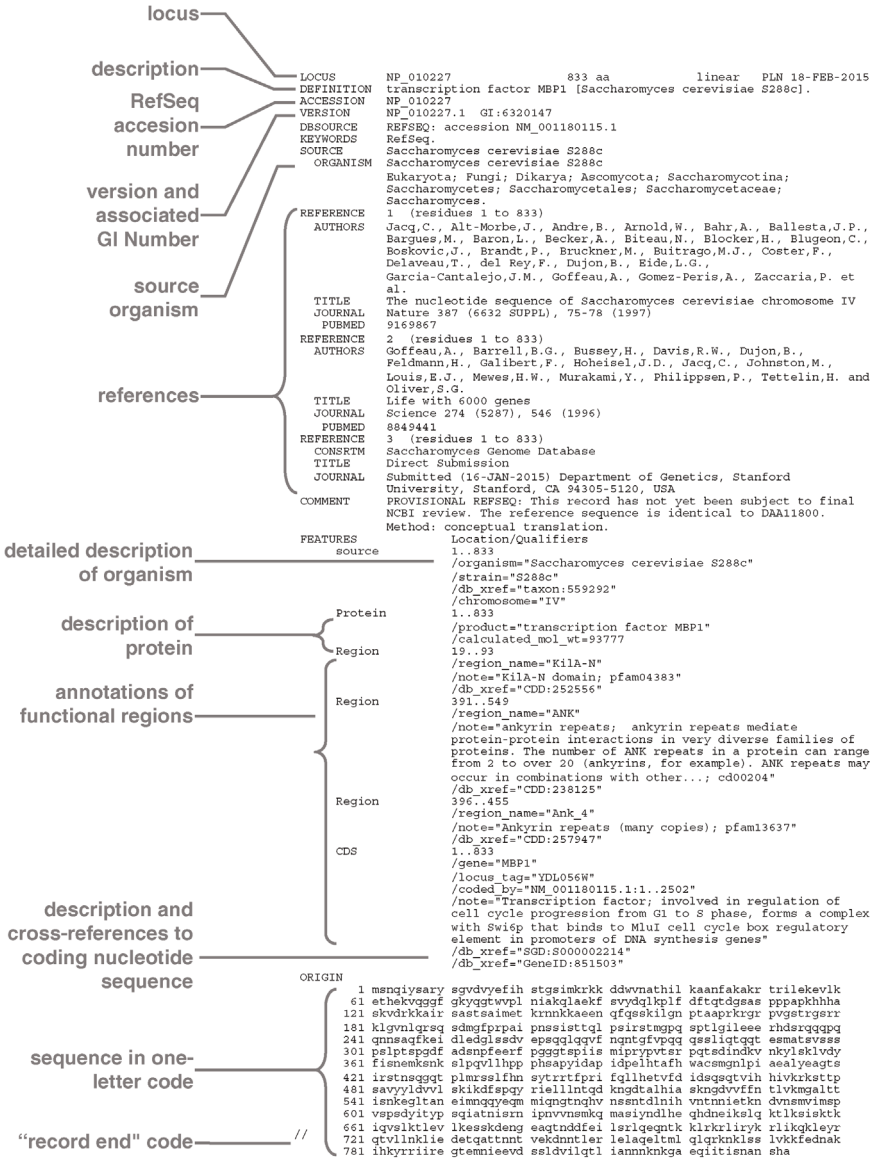


Figure 33.2 "Anatomy" of a GenBank flat-file formatted database entry. Protein sequence of the Mbp1 protein. Elements of the entry are annotated.

only two elements: a "header" line, prefixed with a single ">" character, and any number of lines containing the actual sequence data in one-letter code, until the end-of-file. Often – but not mandatory – lines are 80 characters in length, a vestige of the size-limits of punch-cards for data transfer. The format is compact, efficient, readable, and easy to open and edit in any word-processor. Virtually all analysis programs accept input in FASTA format.

```
>ref|NP_010227.1| transcription factor MBP1 [Saccharomyces cerevisiae S288c]
MSNQIYSARYSGVDVYEFIHSTGSIMKRKKDDWVNATHILKAANFAKAKRTRILEKEVLKETHEKVQGGF
GKYQGTWVPLNIAKQLAEKFSVYDQLKPLFDFTQTDGSASPPPAPKHHHASKVDRKKAIRSASTSAIMET
KRNNKKAEENQFQSSKILGNPTAAPRKRGRPVGSTRGSRRKLGVNLQRSQSDMGFPRPAIPNSSISTTQL
PSIRSTMGPQSPTLGILEEERHDSRQQQPQQNNSAQFKEIDLEDGLSSDVEPSQQLQQVFNQNTGFVPQQ
QSSLIQTQQTESMATSVSSSPSLPTSPGDFADSNPFEERFPGGGTSPIISMIPRYPVTSRPQTSDINDKV
NKYLSKLVDYFISNEMKSNKSLPQVLLHPPPHSAPYIDAPIDPELHTAFHWACSMGNLPIAEALYEAGTS
IRSTNSQGQTPLMRSSLFHNSYTRRTFPRIFQLLHETVFDIDSQSQTVIHHIVKRKSTTPSAVYYLDVVL
SKIKDFSPQYRIELLLNTQDKNGDTALHIASKNGDVVFFNTLVKMGALTTISNKEGLTANEIMNQQYEQM
MIQNGTNQHVNSSNTDLNIHVNTNNIETKNDVNSMVIMSPVSPSDYITYPSQIATNISRNIPNVVNSMKQ
MASIYNDLHEQHDNEIKSLQKTLKSISKTKIQVSLKTLEVLKESSKDENGEAQTNDDFEILSRLQEQNTK
KLRKRLIRYKRLIKQKLEYRQTVLLNKLIEDETQATTNNTVEKDNNTLERLELAQELTMLQLQRKNKLSS
LVKKFEDNAKIHKYRRIIREGTEMNIEEVDSSLDVILQTLIANNNKNKGAEQIITISNANSHA
```

Figure 33.3 GenBank sequence of the Mbp1 protein of baker's yeast in FASTA format. A subset of the information of the GenBank record's header is given here: RefSeq ID and organism description. The sequence data is represented in one-letter code, with line-breaks after 70 characters. There is no explicit end-of-sequence character.

33.3.3 Nucleotide Sequence Management in the Laboratory

Dedicated commercial and open-source solutions exist for managing in-house sequences in the molecular biology laboratory, a web search will identify them readily. Many laboratories simply store sequences in text files on their personal computers. Of course, a robust backup-scheme is critical in all cases.

To edit, translate, and annotate sequence, ApE (A plasmid Editor, http://biologylabs.utah.edu/jorgensen/wayned/ape/) is a popular, free tool for the most common computer platforms.

33.4 Sequence Analysis on the Web

Just as with the databases, finding the right source for analysis services is less a question of what is possible, but how to choose among the many alternatives. The annotated *NAR* issues and the searchable directory of Bioinformatics.ca have been mentioned above; an integrated package of basic analysis function is provided by the EMBOSS package.

33.4.1 EMBOSS

The European Molecular Biology laboratory Open Software Suite (EMBOSS) is a collection of programs for basic sequence analysis tasks (http://emboss.sourceforge.net/). They are generally freely available as services via dedicated Web servers, a simple Google search for "emboss explorer" will find access points (Table 33.3).

The EMBOSS package is open-source software and can easily be installed on a laboratory's local computers. Versions also exist within Biopython; thus the construction of automated workflows is possible.

Table 33.3 Selection of applications of the EMBOSS analysis package.

Application name	Function
trimmest	Remove poly-A ends from EST sequences
einverted	Find DNA inverted repeats
eprimer3	Select PCR primers and hybridising oligonucleotides
geecee	Compute G/C contents of nucleotide sequences
revseq	Reverse complement sequences
remap	Plot a nucleotide sequence with restriction endonuclease sites and translation
banana	Sequence dependent twist and curvature of B-DNA
transeq	Translate nucleotide sequences
pepstats	Peptide statistics: molecular mass, extinction coefficient, and so on
iep	Compute isoelectric point of a peptide
pepcoil	Predict coiled-coil regions
sigcleave	Predict secretion signal peptidase sites
fuzzpro	Discover patterns in protein sequences
pepwheel	Plot peptides as helical wheels to emphasize amphipathic moment
dotmatcher	Dot-plot sequence comparison
needle	Needleman–Wunsch global-optimal pairwise sequence alignment
water	Smith–Waterman local pairwise sequence alignment

```
PEPSTATS of Mbp1 from 1 to 833

Molecular weight = 93907.64             Residues = 833
Average Residue Weight  = 112.734       Charge   = 26.0
Isoelectric Point = 9.8128
A280 Molar Extinction Coefficient  = 43950
A280 Extinction Coefficient 1mg/ml = 0.47
Improbability of expression in inclusion bodies = 0.896

Residue     Number      Mole%
A = Ala     43          5.162
C = Cys     1           0.120
D = Asp     39          4.682
E = Glu     49          5.882
F = Phe     25          3.001
[...]
```

Figure 33.4 Partial output of the PEPSTATS routine of the EMBOSS package, for the amino acid sequence of yeast Mbp1. The molecular weight and coefficient of extinction are properties that are frequently needed in a laboratory.

33.5 Sequence Composition

Several sequence features can be determined directly from the additive properties of its molecular building blocks. This includes composition – the molar ratios of the constituent amino acids – molecular mass, coefficient of extinction, antigenicity, and isoelectric point (which, however, may be shifted due to the variation of pK values in a folded protein) (Figure 33.4).

33.6 Sequence Patterns

Transcription Maps of the Human Genome, Section 36.3.5

As genome sequencing has become readily accessible, the problem of annotation of raw nucleotide data has become increasingly important. Even the distinction between transcribed and untranscribed sequence, between information contained in the translated polypeptide, and regulatory genomic information is challenging. Untranslated sequences may contain functional sites of promoters, terminators, enhancers, isolators, and much more. Such functional sites are mostly associated with the regulation of gene expression – the genome after all encodes both structural and sequential information – but also with the regulation of replication. For the human genome, the ENCODE project (http://genome.ucsc.edu/ENCODE/) is a large-scale effort to annotate the entire genome and provide a list of all of its functional elements. At least for human genes, functional motifs have been experimentally determined and the annotations are available on the Web (Figure 33.5).

Genome-level regulatory information is determined both by sequence and by context. The sequence pattern encodes the functional potential, where and when it is expressed depends on the level of nucleosome compaction, epigenetic modifications, occupancy of activation and silencing sites, and much more. While ENCODE is a fundamentally experimental approach, bioinformatics methods complement the experiments. In principle we can distinguish between analysis by signal, that is, pattern analysis, and analysis by contents, that is, sequence comparison.

Pattern analysis determines the presence or absence of a defined sequence pattern, which may be continuous and exact, discontinues with fixed or varying gap-lengths, and/or may contain ambiguities. We can distinguish between deterministic analysis and probabilistic analysis. *Deterministic* analysis gives a yes/no answer: a pattern is either present or it is not. The classical paradigm is the analysis of restriction endonuclease recognition sites in nucleotide sequence. Algorithms are fast and well understood. Programming languages support such searches through *regular expression* functions. For protein sequences, functional patterns have been compiled in the PROSITE database of domains, families, and functional sites (http://prosite.expasy.org/), which can be used to scan for functional motifs given a Uniprot accession number or an uploaded sequence.

Motifs should be sensitive and specific, that is, define a short pattern that occurs in all sequences that share a particular function and is absent in sequences that do not.

An example for a sequence motif is the pattern:

L-x(6)-L-x(6)-L-x(6)-L

which is Prosite Motif PS00029, a leucine zipper. The notation describes four leucine residues "L", separated by six unspecified residues. Such a "zipper" can be found in a homologue to

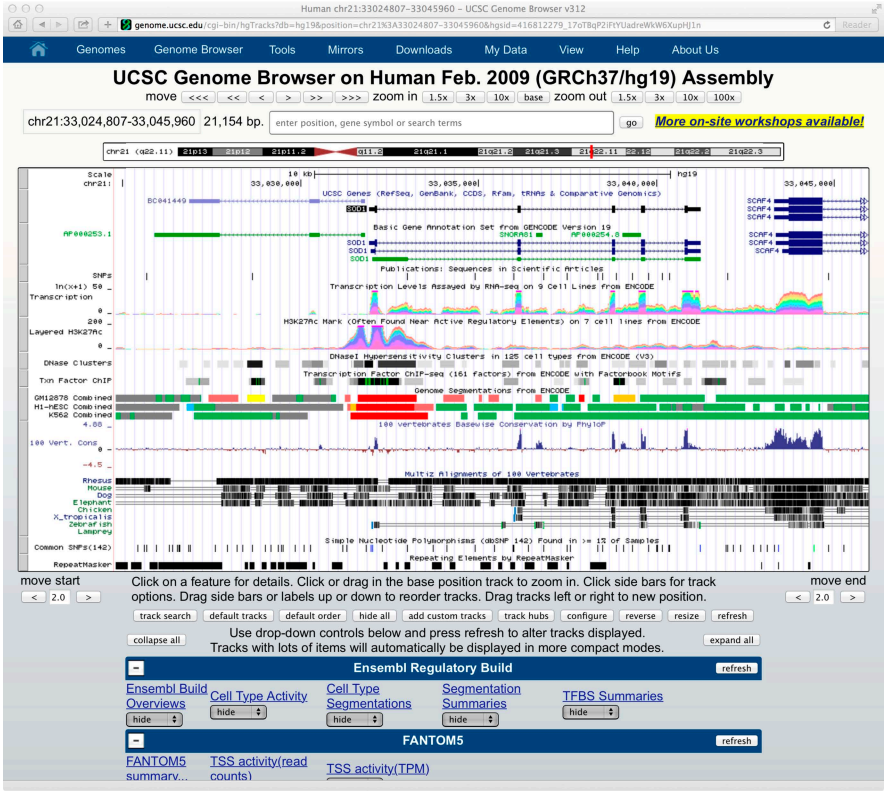


Figure 33.5 ENCODE project annotation patterns overlaid on a human genome map (cf. http://genome.ucsc.edu). The grey bars show experimentally determined ChIP-seq and DNAse sensitivity clusters in the vicinity of the human SOD1 gene. This image indicates the very high level of annotation that is currently available via simple Web queries.

Mbp1, the yeast Swi4 protein with the sequence:

848 – **L**espssl**L**piqmsp**L**gkyskp**L**sqqink**L**ntkvss**L** – **883**

This zipper has not been annotated in the feature table of the Swi4 GenBank entry; this is an example of how such feature tables are helpful for orientation, but are necessarily incomplete.

Each Prosite motif is accompanied by an expert-curated description of its biological implications.

More complex sequence patterns are commonly analyzed through *probabilistic* pattern analysis that aims to quantify the probability that a pattern is present rather than giving only a yes/no answer. This is frequently computed via a position-specific scoring matrix (PSSM), or profile, in which each position of the sequence is associated with a vector of possible states it can take, and the associated probabilities. Such a matrix is a computational tool that quantifies how well a particular sequence matches a given sequence motif. The algorithm simply adds the profile-values for the observed character in each position. If these are properly scaled, the result can be interpreted as a probability that the observed sequence was part of the set that the weighting matrix was based on. Such sequence propensities can be displayed in a so-called *sequence-logo* which represents the information content of every position and its observed states. This corresponds to our biological intuition that conserved positions should be considered to be more important than variable regions (Figure 33.6).

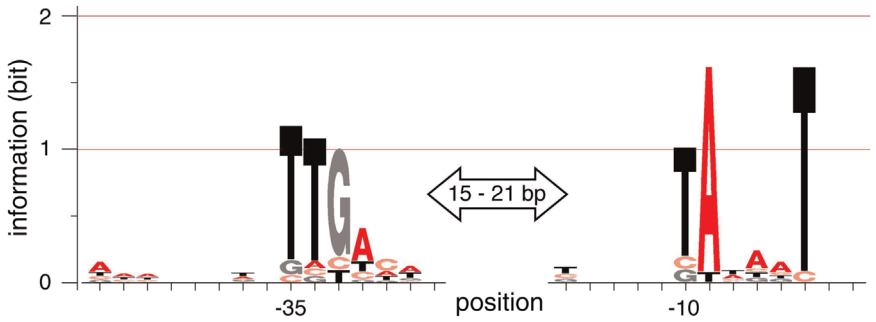


Figure 33.6 Sequence logo, computed from 21 *Escherichia coli* promoter sequences. The height of the character-stack at every position corresponds to the information-theoretical information, in "bits." This corresponds to the weight of a position-specific scoring matrix. The size of the individual letters correspond to their frequency in a position. These logos were pioneered by Tom Schneider, and can be easily generated online from multiple alignments (http://weblogo.berkeley.edu/).

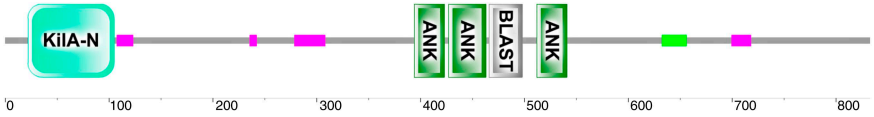


Figure 33.7 SMART server domain and pattern annotation of the yeast Mbp1 protein. Features are mapped to the sequence and detailed annotations can be studied further. In this case we see a KilA-N family DNA binding domain and a series of ankyrin protein–protein interaction modules annotated by comparison to the Pfam database of protein domain families, and an ankyrin domain annotated by sequence similarity to a known domain (BLAST). Other features such as a coiled-coil dimerization domain (green) and unstructured, low-complexity regions (pink) are determined by first principles from the target sequence itself. Other features (not shown) include acetylation and phosphorylation sites, determined from sequence patterns. These local patterns alone are generally necessary but are not sufficient to make a prediction. Whether they are actually functional depends on their local context in the native, folded protein.

Probabilistic pattern descriptions for annotation and classification are the domain of machine learning in computer science. PSSMs make the implicit assumption that the contributions from each position are independent, but many more powerful tools have been developed over the last two decades that can capture higher-order preferences that may be subtly distributed over the entire pattern. Methods include Hidden Markov Models, Neural Networks, Support Vector Machines, Random Forests, and more. They can be implemented, for example, from program packages for the **R** programming language (https://cran.r-project.org/web/views/MachineLearning.html) and excellent online introductory tutorials are available (http://www-bcf.usc.edu/~gareth/ISL/). The common theme is that the high-dimensional features of objects (such as sequences) with known properties are represented in some consistent way; this is called a training set. Then objects with unknown properties – the test set – are queried to determine which subsets of the training set they are similar to. If the similarity is high, the algorithm can conclude that the new object should share annotations with the annotated subset. This strategy is common to all "supervised" machine-learning approaches, they merely differ in the details of how the information is represented but often lead to results of comparable quality.

This leads to very general pattern matching procedures that can be used for sequence annotation, for example, at the SMART database at the European Molecular Biology Laboratory in Heidelberg (http://smart.embl-heidelberg.de/) (Figure 33.7).

33.6.1 Transcription Factor Binding Sites

One example of the role of pattern recognition for sequence analysis is the annotation of regulatory sequences. The sequence patterns for experimentally validated transcription factor binding sites are frequently redundant, and present in the genome much more frequently than we would expect for functional reasons. Which of these sites are actually functional? Databases of validated binding sites such as JASPAR (http://jaspar.genereg.net/) can help answer this question. JASPAR allows us to select sets of curated transcription factor binding profiles and predict binding sites in uploaded sequences (Figure 33.8).

If only genomic sequences of co-regulated genes are known, but the binding sites and factors have not been determined, motif discovery algorithms can be applied. The principles for several successful algorithms are similar: given a set of nucleotide sequences that are known to contain binding sites somewhere, mask trivially shared sequences and then test all possible sequence factors of a given size to find those that are significantly overrepresented. Excellent results have been achieved with the programs Weeder (http://159.149.160.51/modtools/), Meme/MAST (http://meme.nbcr.net/meme/), the Gibbs Motif Sampler (http://ccmbweb.ccv.brown.edu/gibbs/gibbs.html), and most recently XXmotif (http://xxmotif.genzentrum.lmu.de/). A server that integrates several approaches by running them independently and clustering the results is WebMOTIF (http://fraenkel.mit.edu/webmotifs-md-programs.html).

Beyond merely identifying statistically overrepresented patterns, their functional relevance can be evaluated in the light of biological knowledge. Criteria that support functional relevance include:

- enrichment of motifs in the upstream untranscribed regions of co-regulated genes, relative to the regulatory regions of randomly chosen genes;
- conservation in syntenic genomic regions of related organisms;
- higher propensity to appear together with other validated binding sites of regulatory motifs.

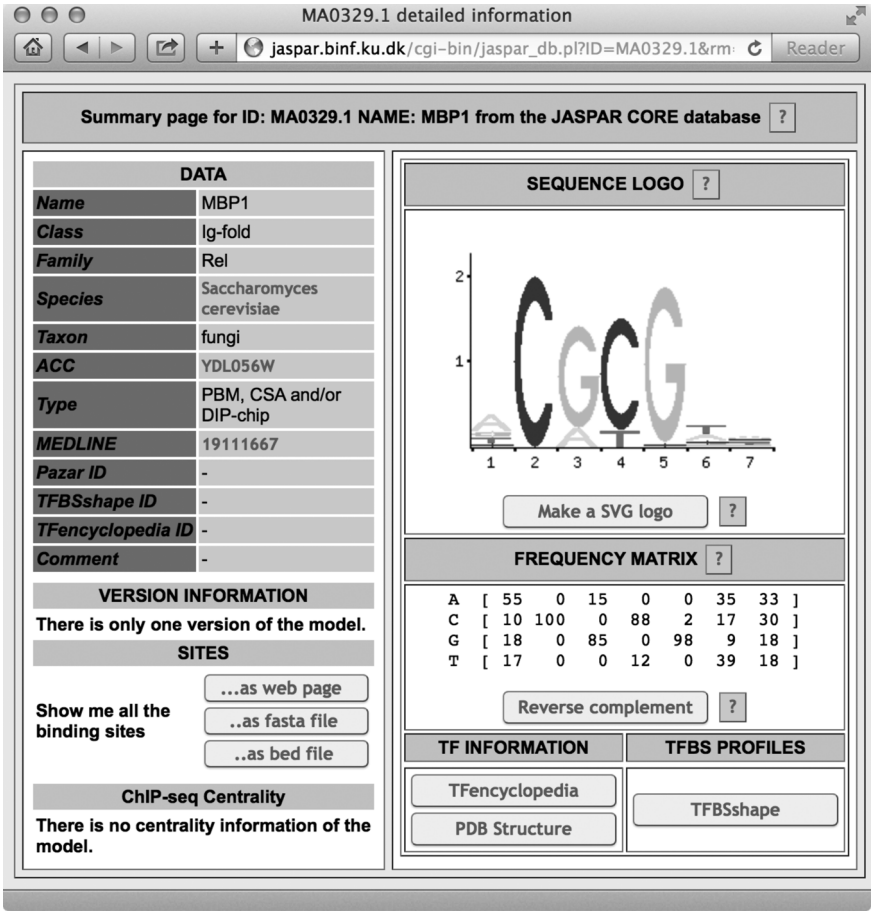


Figure 33.8 Information window for the binding profile of yeast Mbp1, curated in the JASPAR database. The binding site propensity matrix is shown as a scoring matrix. Links lead to additional information resources. It is easy to appreciate that the site is at its core only a tetranucleotide, and the biological function of the transcription factor is established through contextual binding in concert with other factors.

Fundamental to this type of analysis is the hypothesis that co-occurring motifs indicate shared function: we call this "guilt by association." Obviously, there may also be other reasons for the association of recurring patterns, such as segmental duplications. Importantly, highly repetitive sequences can appear as overrepresented patterns, which is technically correct but biologically a meaningless result. Such sequences can be identified with the program RepeatMasker (http://www.repeatmasker.org/) and excluded from analysis.

One application domain is the annotation of motifs in "peaks" – frequently identified genomic regions – of ChIP-seq or ChIP-chip experiments. For example, the Bioconductor ChIPpeakAnno package provides functions to retrieve collated genome features and additional annotations such as identifying multiple-transcription factor loci (MTLs) and functional annotations from the Gene Ontology database (GO, http://geneontology.org/).

Analysis of Epigenetic Modifications, Section 31.7

33.6.2 Identification of Coding Regions

Genomes consist in their majority of untranslated sequences, thus the identification of coding regions is one of the key challenges of newly sequenced genomic DNA. This too is a task for pattern recognition, albeit the issue here is less one of recognizing particular signals such as splice donor and acceptor signals. These are important, but their predictive value is determined by an array of more delocalized properties: G/C contents, relative di-, tri-, and hexanucleotide frequencies, absence of stop-codons, degree of conservation in related genomes, and other measures. RNA-seq as an experimental method to discover transcribed regions has become

accessible and increasingly important to complement the purely computational approaches. Once again, machine learning methods play a big role to integrate large numbers of features to evaluate evidence.

Early successes were primarily based on Hidden Markov Models. GENSCAN is a popular example for eukaryotes (http://genes.mit.edu/GENSCAN.html) while Glimmer is an application that has annotated hundreds of prokaryotic and viral genomes (http://ccb.jhu.edu/software/glimmer/). Newer approaches, however, make extensive use of EST and RNAseq data and this gives improved performance for situations that are difficult for purely computational approaches: short exons, alternative splicing, overlapping genes, non-canonical splice sites, and RNA editing. A program that facilitates the integration of various types of evidence is JIGSAW (https://www.cbcb.umd.edu/software/jigsaw/).

Recent interest has also focused on microRNA prediction and MiPRED (http://www.bioinf.seu.edu.cn/miRNA/) provides a tool with satisfactory performance.

This type of analysis scales well to whole-genome annotations, but actually using the additional information that is available through the large number of previously sequenced genomes requires a bit more thought. For prokaryotic genome annotation, the RAST annotation engine (Rapid Annotation using Subsystem Technology, http://rast.nmpdr.org/) accesses the SEED collection of functionally related protein families to annotate genes by comparison to the collection. Once the genome is made public, the annotated proteins are added to the collection, which thus provides a constantly improved, growing body of functional annotation information.

33.6.3 Protein Localization

Secretory proteins have a characteristic N-terminal signal sequence that is cleaved after translocation through the inner membrane. The information about the precise cleavage site is distributed along the sequence, which has several characteristic features (Figure 33.9).

Predicting these cleavage sites has been one of the earliest successful applications of neural networks to molecular biology and the current version of the SignalP algorithm performs well enough to make it suitable for automated, whole genome annotation (http://www.cbs.dtu.dk/services/SignalP/), addressing the non-trivial challenge of distinguishing secretion signals from transmembrane helices. It achieves a Matthews correlation coefficient (a number characterizing the relationship between sensitivity and specificity of a procedure) of better than 0.8, and a best-in-class performance of all currently available alternatives.

Similar algorithms are suitable for the prediction of protein localization, transmembrane helices, domains, and so on – one server that integrates a number of such predictions is PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/), alternatives include PSORT for localization prediction (http://www.psort.org/) and TMHMM (http://www.cbs.dtu.dk/services/TMHMM) and TOPCONS (http://topcons.cbr.su.se/) for transmembrane helices.

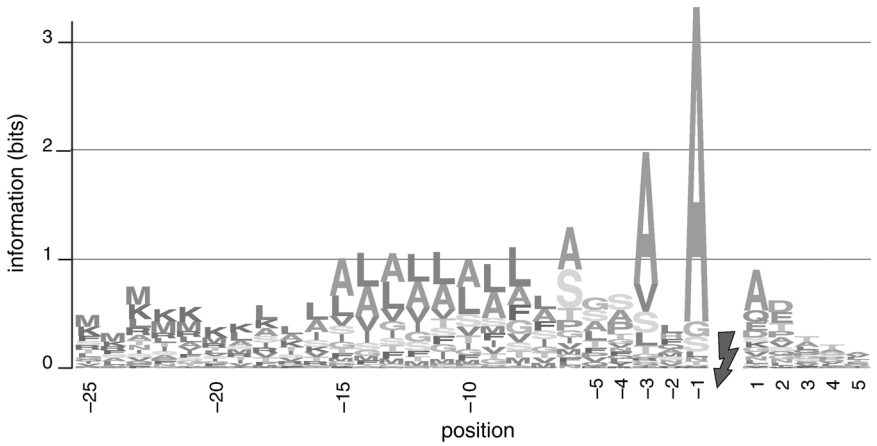


Figure 33.9 Sequence logo for the N-terminus of Gram-negative bacterial secreted proteins, including the signal peptide. There is a clear preference for a positively charged N-terminus, a hydrophobic central stretch with amino acids that have high helical propensity, and a preference for small amino acids in positions −3 and −1 of the cleavage site. Downstream of the cleavage site, hydrophilic residues predominate.

33.7 Homology

Pattern-based methods can contribute important basic information to sequence analysis, for practical purposes, however, homology based methods are far more important. The basic procedure is to construct an optimal alignment between two sequences and to evaluate it. If the alignment suggests that the sequences diverged from a common ancestor, one can deduce domain arrangements, phylogenetic history, important functional residues, and much more. The wealth of information is based on the empirical observation that related sequences usually have related function and virtually always have similar structure, and that important functions are associated with the evolutionary conservation of residue motifs.

33.7.1 Identity, Similarity, Homology

Identity, similarity, and homology are terms with a precise meaning but they are often used in an imprecise way.

Identity is the percentage of residues that are identical in an alignment.

The quantification of similarity is difficult (Figure 33.10). It requires a measure of similarity between pairs of amino acids – for example, how similar one would consider a valine to be to a threonine residue or to a leucine. One can base such a measure on biophysical similarity, on the number of required nucleotide changes to get from one to the other, or on other considerations. There is no accepted general way to define similarity, however, and therefore the term should not be used in a quantitative sense.

Homology on the other hand means that two proteins have diverged from a shared ancestor. It has an exact meaning and describes a quality, not a quantity. Two proteins can be homologous, or not. To say that two genes share "50% homology" abuses the terminology. Homology does not necessarily have to imply identity or similarity, since homologous sequences can be diverged to a degree that they have lost all recognizable similarity. However, if the degree of similarity is high, shared ancestry is a much more likely explanation than random chance.

In the following, we concentrate on amino acid sequences. Since nucleotide variability is generally much higher than amino acid variability, it is usually not very meaningful to compare the untranslated nucleotide sequences of coding regions. Obviously, this does not apply to sequences in which the nucleotides themselves are functionally conserved, such as in promotor-, t-RNA sequences. For these cases the methods of amino acid sequence comparison can be readily generalized – with the exception perhaps that helical stems can have conserved H-bonding patterns despite sequence change and thus may need to be separately considered.

We further assume that sequence similarity is an additive property of independent sequence positions. This is a crude approximation that could be improved upon if additional information about amino acid interactions were available. In practice this is, however, usually not the case and assuming independence is the best we can do. Nevertheless, there is potential for improvement of alignment algorithms that could include secondary structure propensities, conservation patterns, solvent accessibility, or the precise position of gaps in the alignment – albeit at the cost of much added algorithmic complexity.

Similarity scores of individual amino acid pairs are commonly represented in a "scoring matrix." All results that use such a matrix for alignment obviously depend on what these values are. The matrix is a tool to quantify how well an alignment reflects a particular model of similarity. If we construct the matrix to represent evolutionary relationships, an optimal alignment – that is, an alignment that gives the highest possible aggregate score for this matrix – will be suitable to evaluate homology. We consider two proteins as likely homologues if the alignment score is higher than what we would expect from two randomly chosen sequences.

The matrix that is most commonly chosen for this task is BLOSUM 62, developed by Jorja and Steve Henikoff; it captures exchange probabilities from blocks of aligned, un-gapped sequences at minimally 62% sequence dissimilarity (Figure 33.11). This approach reflects the fact that amino acids from gapped alignments are in structurally incomparable regions, and their alignment is meaningless in the first place.

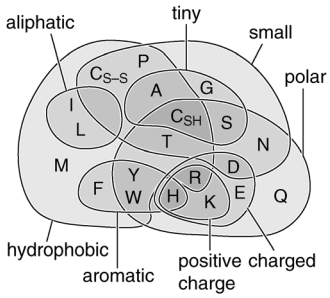


Figure 33.10 An overview of the biophysical properties of amino acids as a Venn diagram. Amino acids are given in the one-letter code, cysteine appears twice since the properties of the cysteine side chain's free thiol (C_{SH}) and the disulfide bonded cystine molecule (C_{S-S}) are very different. Individual properties can be accurately quantified, for example, "hydrophobicity" can be measured as the free energy of transfer from a water-phase to octanol, "size" can correspond to the volume of the solvent-accessible surface of a residue. However, their relative weight is arbitrary and such measures cannot be combined into a single number in a rigorous way. Which measure is the most appropriate depends on the question. To quantify similarity for the purpose of sequence alignment, measures based on the ability of one amino acid to substitute for another in natural sequences have proven most successful. (Diagram after and adapted from Taylor, W. R. (1986) The classification of amino acid conservation. *J. Theor. Biol.* **119** (2), 205–218.)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**4**	−1	−2	−2	0	−1	−1	0	−2	−1	−1	−1	−1	−2	−1	1	0	−3	−2	0
R		**5**	0	−2	−3	1	0	−2	0	−3	−2	2	−1	−3	−2	−1	−1	−3	−2	−3
N			**6**	1	−3	0	0	0	1	−3	−3	0	−2	−3	−2	1	0	−4	−2	−3
D				**6**	−3	0	2	−1	−1	−3	−4	−1	−3	−3	−1	0	−1	−4	−3	−3
C					**9**	−3	−4	−3	−3	−1	−1	−3	−1	−2	−3	−1	−1	−2	−2	−1
Q						**5**	2	−2	0	−3	−2	1	0	−3	−1	0	−1	−2	−1	−2
E							**5**	−2	0	−3	−3	1	−2	−3	−1	0	−1	−3	−2	−2
G								**6**	−2	−4	−4	−2	−3	−3	−2	0	−2	−2	−3	−3
H									**8**	−3	−3	−1	−2	−1	−2	−1	−2	−2	2	−3
I										**4**	2	−3	1	0	−3	−2	−1	−3	−1	3
L											**4**	−2	2	0	−3	−2	−1	−2	−1	1
K												**5**	−1	−3	−1	0	−1	−3	−2	−2
M													**5**	0	−2	−1	−1	−1	−1	1
F														**6**	−4	−2	−2	1	3	−1
P															**7**	−1	−1	−4	−3	−2
S																**4**	1	−3	−2	−2
T																	**5**	−2	−2	0
W																		**11**	2	−3
Y																			**7**	−1
V																				**4**

Figure 33.11 The BLOSUM62 mutation data matrix. Amino acids are given in the one-letter code. Positive values indicate high probability of exchange in related sequences and increase the alignment score, negative values decrease the score and indicate that such pairs would be more likely in an alignment of unrelated sequences. An F → Y pair is highlighted, its value is +3. Values on the diagonal correspond to the probability of residues to be conserved. For example, a conserved tryptophan (+11) or cysteine (+9) is a much more significant indicator of homology than the more generic alanine, leucine, or serine (+4).

33.7.2 Optimal Sequence Alignment

How can we use the mutation data matrix to determine the correct alignment in which all matched amino acid pairs have descended from the same ancestral residue? The simple answer is: We can't. In general we cannot prove that a particular alignment is "correct," unless we could observe the sequence through its entire evolutionary trajectory. But what we can do is to compute an optimal alignment. To the degree that the mutation data matrix represents evolution, the optimal alignment should be our best guess as to what is correct. And to compute an optimal alignment should be simple – just create all possible alignments, score them, and pick the best.

Unfortunately, it turns out that this cannot be done. Natural sequences do not always conserve their length, we frequently find insertions and deletions. An insertion for one sequence is a deletion for the other, thus we usually use the portmanteau "indel" to describe this. Since indels exist, if we want to generate all possible alignments, we need to consider three options at every position: a match, an insertion, and a deletion. Thus the number of possible alignments for two sequences of 100 amino acids in length is larger than the number of particles in the visible universe. We generally consider computational problems of that size to be intractable.

Fortunately an efficient algorithm to solve this problem was discovered by Saul Needleman and Christian Wunsch in the 1970s. The principle is quite straightforward and relies on the fact that we are treating the aligned amino acid pairs as independent. In this case, assume we want to align two sequences of 100 amino acids optimally. We can simplify this problem by computing an optimal alignment for 99 amino acids, and extending it in the best possible way. But how do we compute that alignment? We can simplify this problem by computing an optimal alignment for 98 amino acids, and extending it in the best possible way. And so on . . . until we need to compute the optimal alignment for just one amino acid pair. This score we can simply retrieve from the scoring matrix. In this way, building the alignment from the bottom up, we can construct the alignment in a *recursive* procedure.

The problem is that the recursive procedure is still not efficient enough to be tractable. But considering how it is computed, Needleman and Wunsch noticed that the same partial alignments needed to be computed over and over again. That is of course not efficient – it

```
#=======================================
# Aligned_sequences: 2
# 1: Swi4
# 2: Mbp1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1147
# Identity:     266/1147 (23.2%)
# Similarity:   414/1147 (36.1%)
# Gaps:         368/1147 (32.1%)
# Score: 640.0
#=======================================

    1 MPFDVLISNQKDNTNHQNITPISKSVLLAPHSNHPVIEIATYSETDVYEC     50
                                       .||.  |..|.||..||||
    1 ------------------------------MSNQ--IYSARYSGVDVYE-     17

   51 YIRGFETKIVMRRTKDDWINITQVFKIAQFSKTKRTKILEKESNDMQHEK    100
      :|.  .|..:|:|.||||:|.|.:.|.|.|:|.|||:|||||.....|||
   18 FIH--STGSIMKRKKDDWVNATHILKAANFAKAKRTRILEKEVLKETHEK     65

  101 VQGGYGRFQGTWIPLDSAKFLVNKYEIIDPVVNSILTFQFDPNNPPPKRS    150
      ||||:|::||||:||:.||.|..|:.:.|.:.......|.|.:..||...
   66 VQGGFGKYQGTWVPLNIAKQLAEKFSVYDQLKPLFDFTQTDGSASPPPAP    115

  151 KNSILRKTSPGTKITSPSSYNKTPRKKN-SSSSTSA--TTTAANKKGKKN    197
      |:                ...:|..|||. .|:||||  .|...|||.::|
  116 KH---------------HHASKVDRKKAIRSASTSAIMETKRNNKKAEEN    150

      [ ... ]
```

Figure 33.12 Partial output of the program *needle* in the EMBOSS package. The sequences of yeast Mbp1 and its homologue Swi4 have been aligned, the region covering the DNA binding domain is shown. Identical amino acid pairs are marked with a "|" character, more or less similar characters are marked with ":", respectively ".".

is much better to compute partial alignments only once, store them, and reuse them when necessary. This strategy is called memoization or dynamic programming and applying it makes the algorithm quite fast – in fact its computational requirements scale only with the square of the alignment length: 10 000 steps versus 10^{100} for the brute force approach is a rather large difference. Several details need to be clarified like how to quantify indels, and which global optimum is reported in case there are several equally good ones, but this does not change the essence of the approach.

There is no quantitative theory that describes the probability and length of indels through evolutionary change. Empirically we find that indels are less frequent the longer they are, but there is evidence that there may be more than one distinct mechanism that causes them, since log-length versus log-frequency plots of indels from database-scale comparisons show distinct regimes. Still, it is computationally efficient to simply model indels with a constant insertion penalty and an extension penalty that depends on the indel length. The actual values obviously have to harmonize with the pair-scores in the mutation data matrix; typical valves assure that an indel of two residues or so will need to be justified by two extra identities in the alignment to improve the overall score. Default values vary a bit between implementations, for example, the NCBI uses (insertion, extension) penalties of (11, 1), for BLAST alignments (see below), while the EMBOSS package uses (10, 0.5) for optimal sequence alignments (Figure 33.12), both for the BLOSUM62 matrix. Users can adjust the parameters to favor or disfavor indels and perhaps gain some intuition concerning which parts of an alignment are robust and which parts depend critically on the detailed alignment parameters.

The Needleman/Wunsch alignment computes *global* optimal alignments that cover the full length of both sequences. Frequently we are, however, interested in *local* optimal alignments that cover only those parts of a sequence that can be well aligned in the first place. This is especially important if our sequences have very different lengths, or if we want to focus only on individual domains – which is always to be recommended, since multidomain proteins are not necessarily homologous over the entire length. Temple Smith and Michael Waterman proposed a variation of the Needleman/Wunsch algorithm that achieves this. Two modifications are required: first of all, the scoring matrix must be scaled such that it has a negative overall expectation value for pair scores. Otherwise a random extension of an alignment would still improve the score. Secondly, once a partial alignment has a negative score, that score is set to zero. Then the highest scoring partial alignment is sought out; this marks the beginning of the

best local alignment. Its sub-alignments are considered, until their score drops to zero, which marks the end of the best local alignment.

33.7.3 Alignment for Fast Database Searches: BLAST

While the optimal sequence alignment is the best we can create, it requires significant resources to compute on a large scale. And by large scale we mean comparing millions of sequences. In fact the protein section of the NCBI's non-redundant Refseq database consists of on the order of 100 million sequences. Steven Altschul and colleagues published an algorithm in 1990 that is much better suited for the search of homologous sequences in large databases than optimal alignment. The alignments are not guaranteed to be optimal, but the differences are usually negligible. But the NCBI's BLAST (Basic Local Alignment Search Tool, blast.ncbi.nlm.nih.gov/) services are powerful enough to freely process the world's sequence searches each day; BLAST is the most frequently used bioinformatics tool by far.

The algorithm works by breaking a query sequence into small fragments and searching for the occurrence of clusters of fragments in a large index table. Once matches are found, the algorithm attempts to extend them locally without indels. Ungapped seed alignments are evaluated for statistical significance and joined to full-length alignments. The algorithm is fast, because the first step – the look-up of initial matches in a table – is fast and thus the efforts of the program are concentrated in those regions of database sequences where matches are promising, while ignoring the very large sequence space in which no significant alignment would be possible in the first place.

A key feature of this process is the computation of statistically meaningful measures of significance of a match – BLAST calls these *E*-values. An *E*-value computes the probability that an alignment of the same quality for two *unrelated* sequences would appear in a database of the same size. This means the *E*-value is not a measure of whether two sequences are homologous, but a measure of whether two unrelated sequences could have the same score as the value being considered. This depends on the database size – and paradoxically this means that the more sequences are available, the harder it becomes for a given alignment to achieve significance. As a corollary, one needs to choose the database carefully in which a sequence search is performed. It should be the smallest database (RefSeq, usually) and the most stringent subset of organisms that we are interested in. Organism subsets can simply be selected on the BLAST input form.

33.7.4 Profile-Based Sensitive Database Search: PSI-BLAST

What are the options if a BLAST search yields too few homologous sequences? In many cases a profile-based search with the PSI-BLAST algorithm can yield additional results. Profiles average the sequence information of individual hits and thus increase the sensitivity of searches by enhancing the signal-to-noise ratio of alignments. In a sense, profiles are position specific scoring matrices over the length of a sequence, and they allow the search to focus on the structurally and functionally important conserved regions, while down-weighting spurious similarities in highly variable segments of the sequence. PSI-BLAST can simply be selected as a search option on the NCBI's BLAST input form, and initially works just like a BLAST search. After a first search, hits are assembled into a profile and the search process is iterated. Typically, new sequences whose alignments were initially not significant are discovered by searching with the profile, and they can themselves be added to the profile in several rounds, may be until convergence, that is, until no new homologues are found. In this way homologous sequences can be identified, even though pairwise sequence identity is far below significance thresholds.

This is, however, not an entirely automatable process: user discretion is required when adding sequences, since even the addition of a single non-homologous sequence by mistake will lead to *profile corruption* – the profile will pull in more and more of the false-positive match's homologues and lead to completely misleading results. To guard against this, one should evaluate sequence annotations and be wary of sequences that have very different annotated functions or cellular locations. Sequences that cover only small portions of the query may also be suspicious. In case of doubt, it is better to manually exclude a questionable

sequence and observe how its *E*-value changes from iteration to iteration. If the *E*-value gets worse or stays approximately constant, this means that the increase of information in the profile does not agree with the sequence information in the questionable sequence and it is unlikely to be a member of the family. Homologous sequences' *E*-values are expected to increase significantly throughout the iterations.

33.7.5 Homology Threshold

Finally, what is the threshold of similarity at which we may conclude that two sequences are indeed homologous? Homologues may have no recognizable sequence similarity at all, and are sometimes discovered solely from structural similarity, similar function, and similar organization of a protein's active site. Low sequence similarity does not necessarily exclude homology. On the other hand, as a good rule of thumb, there are virtually no sequences with more than 25% identity over the length of a structural domain that are *not* homologous. This value is therefore often used as a cut-off for pairwise comparison.

33.8 Multiple Alignment and Consensus Sequences

In the post-genomic era, the likelihood is very high that database searches will discover more than one homologue to a given target sequence. The additional information that can be derived from such sequences is significant since it yields information about conservation of individual positions that correlates with their structural- and/or functional importance. A multiple sequence alignment can additionally help in defining domain boundaries, evaluating whether sequence annotations are conserved, and may serve as the input for analyzing phylogenetic relationships. However, while optimal pairwise alignments are exactly computable, this approach cannot be extended to multiple alignments for two reasons. First, there are practical limitations to the algorithm – the resource requirements grow exponentially with the number of aligned sequences. More importantly however, the objective function, that is, the score that our algorithm should optimize is hard to define in a biologically meaningful way. Should we maximize column-wise sequence similarity? Should we minimize the number and size of indels, since these are rare evolutionary events? Should we aim to cluster indel sites outside of secondary-structure elements? Or should we ensure that sequence patterns and motifs are conserved as highly as possible? Each of these objectives suggests a different computational strategy; however, they cannot necessarily be simultaneously achieved. Even today, multiple sequence alignment is not a solved problem and promising new algorithms appear almost every year.

Different alignment programs will indeed give different alignments and it is not trivial to choose the best. We can certainly say that the first generation algorithms like CLUSTAL are no longer state-of-the art and definitely should no longer be used. But many modern algorithms show only marginal differences when benchmarked against a curated database of structurally aligned reference domains like BaliBASE (http://www.lbgi.fr/balibase/). Tools that can be highly recommended include T-Coffee, MUSCLE, and MAFFT; these regularly outperform other algorithms in direct comparisons and they are all accessible through a launch page at the EBI (http://www.ebi.ac.uk/Tools/msa/). However, the "winner" by a small margin is often the program ProbCons (http://probcons.stanford.edu/).

In practice one will be well advised to perform alignments with several different programs and compare them carefully to identify those parts of the alignment that are robust and independent of the details of algorithm and parameter (Figure 33.13).

The program Jalview (https://www.jalview.org/) is an excellent tool to organize alignments, edit them – yes, manual editing and improvement is encouraged since biological background knowledge can often improve an automated alignment – then compute alignment features, including a first-look phylogenetic analysis, and manage annotations. Alignments can be imported and exported in various formats and can be efficiently computed on a dedicated server.

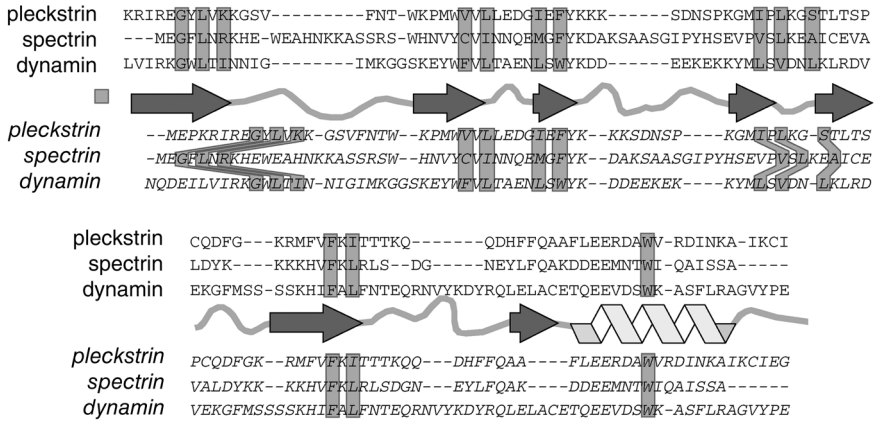


Figure 33.13 Multiple sequence alignment of the N-terminal PH domain of pleckstrin and the spectrin and dynamin PH domain. For these proteins structure coordinates are available and the alignment can be validated with the structural superposition. The top three sequences thus represent the ground truth, the bottom three sequences are excerpts of a larger alignment that was computed using CLUSTAL. Secondary structure elements are shown and the conserved residues of the hydrophobic core shaded. Errors in the CLUSTAL alignment are visible where the shaded columns are not aligned correctly.

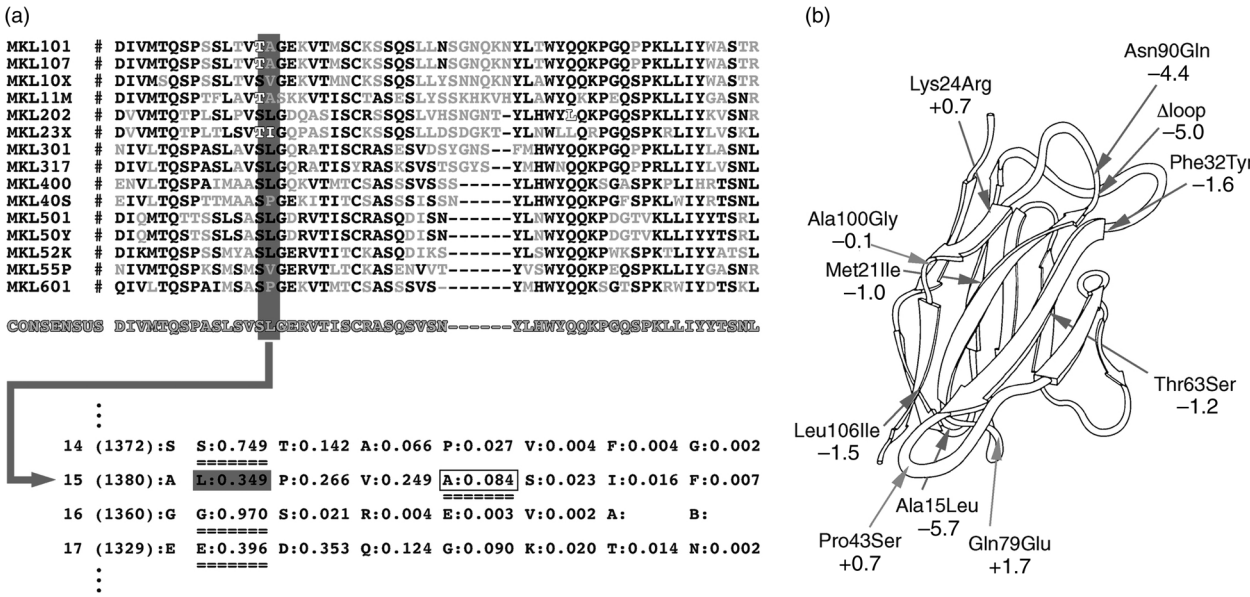


Figure 33.14 Use of a large multiple alignment to engineer protein stability in an immunoglobulin VL variable domain. (a) A sequence alignment is used to determine positional amino acid frequencies. For example, leucine appears four times as often as alanine in position 15 of the sequence. (b) Experimentally determined reversible folding stability changes: A15L stabilizes the domain by 5.7 kJ mol^{-1}. Virtually all predicted mutations could be experimentally verified. Source: Steipe, B. (2004) Consensus-based engineering of protein stability: from intrabodies to thermostable enzymes. *Methods Enzymol.*, **388**, 176–186. With permission, Copyright © 2004 Elsevier Inc.

A remarkable application of multiple alignments is the use of consensus sequences for protein stability prediction (Figure 33.14). It can be shown that changes in amino acids that make a sequence more similar to a family-consensus sequence in general stabilize a protein. Such changes have additive, independent effects to a good approximation.

33.9 Structure Prediction

The accurate prediction of sequence from structure has long been regarded as the holy grail of computational biology. While we have seen impressive results in this respect in the past years, we cannot yet consider it to be a completely solved problem. The quest for structure prediction has, however, lost its importance in the field that has moved away from considerations of individual molecules and towards more comprehensive questions of systems biology.

Since homologous proteins have similar structures, a careful sequence alignment can be used as the basis for homology based structure prediction. This simply consists in replacing the coordinates of aligned side chains. For proteins with about 90% sequence identity, homology models are about as accurate as experimentally determined structures. Unfortunately, regions that contain indels cannot be accurately modeled and even elaborate applications of force-field based energy minimization have not been consistently shown to actually improve the modeled structure. The positive side of this is that it keeps homology modeling simple and for most purposes Web server based modeling, for example, at the SwissModel server (http://swissmodel.expasy.org/), is perfectly adequate. The critical step of homology modeling is not the generation of 3D coordinates, but producing an accurate alignment between target and template. One should aim to produce the best possible alignment from a number of selected homologues, perhaps manually edited with Jalview as the basis of the model. Such alignments can be uploaded to the server and are likely superior to the automated procedures offered by the server itself.

Automated *ab initio* predictions have been successful in many cases with the Rosetta program, pioneered by David Baker. The algorithms are available through the online ROBETTA server (http://robetta.bakerlab.org/) or from the open-source collaboration of Rosetta commons (https://www.rosettacommons.org/). Highlights of successful predictions include building models for crystallographic phasing, and successful *de novo* enzyme design.

33.10 Outlook

Entering the post-genomic era of molecular and cellular biology has changed the field of sequence analysis profoundly. Procedures are easily available on the Internet and can be executed on standard workstations or on free public servers. Genome browsers (http://genome.ucsc.edu/) deliver the annotated maps of whole genomes at the click of a mouse to researchers around the globe. Problems of data integration are being addressed and solved step by step. At the same time, novel problems become apparent, especially regarding mining large datasets for information, visualizing large, high-dimensional relationships, filtering relevant information from the abundance of available resources, and pursuing the complex relationship between genotype and phenotype.

We are still far from routinely making confident predictions about cellular processes and their dynamics. However, modern methods of sequence analysis provide a multitude of views on the function of individual components and the stage is set to improve our understanding of how to integrate these views to larger systems. Obviously, computational methods will continue to grow in importance as we pursue a deeper understanding of life.

34 Analysis of Promoter Strength and Nascent RNA Synthesis

Renate Voit

University of Applied Sciences Bonn-Rhein-Sieg, Department of Natural Sciences, von-Liebig-Straße 20, 53359 Rheinbach, Germany

Maintenance of the basal physiological functions of a cell depends on tissue-specific regulation of gene expression at a particular time. A complex network of cellular pathways and factors control transcription, the first step in gene expression leading to a cell type-specific transcriptome. The nucleus has a highly ordered structure with the genome packaged into more or less tightly compacted chromatin. Since highly compacted chromatin inhibits access of proteins involved in gene expression, transcribed genes are usually located in less compacted areas, the euchromatin, permitting transcription factors to access promoters and adjacent regulatory cis-sequence elements. There are several steps at which gene expression can be regulated: Long-range first, promoter-proximal and promoter-distal regions including locus control regions (LCRs).

Many questions concerning gene expression require analysis of the corresponding RNA produced from the gene locus. In addition to determination of the amount of transcript, which usually reflects the rate of gene expression and strength of the promoter, analysis may consist of mapping of the 5′ and 3′ ends of the transcription unit and mapping of alternative splice variants. Analysis of promoters can be used in one of two ways. The cis-active sequence elements can be altered with *in vitro* mutagenesis to elucidate the effects of these mutations on the rate of mRNA synthesis. Alternatively, the wild-type promoter can be used to determine how the amount of mRNA transcripts from a chosen gene varies under different physiological conditions. Experimental solutions for both types of question are offered by *in vitro* transcription in cell-free systems or the analysis of cloned genes *in vivo* after transfection and expression in cultured mammalian cells. In addition, epigenetic factors such as histone modifications and DNA methylation alter gene expression levels. Experimental approaches related to these topics have been described in detail in other chapters of this book. Histone modifications at specific gene loci can be mapped most effectively with the chromatin immunoprecipitation (ChIP) assay.

34.1 Methods for the Analysis of RNA Transcripts

34.1.1 Overview

To understand the regulation of a particular gene, it is important to know the amount of RNA produced by that gene. The quantification of the gene of interest delivers an important indication of the strength of gene expression. Several methods are available to determine the amount of RNA produced by a gene. A prerequisite for the successful execution of all the described methods is excellent quality of the RNA employed. To isolate intact RNA molecules from cells or tissues, contamination with ribonucleases and unspecific degradation during extraction and isolation must be avoided.

Isolation of Intact RNA, Section 26.6

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

The methods described below for the analysis of RNAs are based on the hybridization of nucleic acids. Since complementary RNA and single-stranded DNA form a very stable hybrid, the resulting RNA-RNA or RNA-DNA hybrid molecules can be qualitatively and quantitatively detected directly (Northern blot, dot blot) or after treatment with single-strand specific nucleases (nuclease S1, ribonuclease A and T1). With the aid of nuclease S1 analysis and ribonuclease protection assays (RPAs), it is possible to quantify the amount of an RNA of interest and map its introns, and the 5′ and 3′ ends on its gene. The primer extension assay uses oligonucleotides complementary to the RNA of interest that are hybridized to the isolated RNA. The resulting hybrids are extended to their 5′ ends with reverse transcriptase (RT) to generate cDNAs. This technique allows the amount of gene-specific RNA to be measured and the 5′ end of the RNA to be determined, even across intron-containing regions. With the fourth technique, Northern blot hybridization, the amount and the mRNAs absolute size can be determined but not the precise locations of its ends. A variant of the Northern blot, dot blot analysis, only serves to quantify the amount of RNA. The most modern method for quantification is quantitative RT-PCR, a method that involves first transcribing the mRNA into cDNA with RT, followed by analysis of the amount of cDNA with the aid of gene-specific primers.

Northern Blot, Section 27.4.4

Dot-blot, Section 27.4.5

Polymerase Chain Reaction (PCR), Chapter 29

34.1.2 Nuclease S1 Analysis of RNA

This widely used procedure, published by Berk and Sharp back in 1977, employs the enzyme nuclease S1, a single-strand specific ribo- and deoxyribonuclease from the fungus *Aspergillus oryzae*. It is unusual in that it only hydrolyzes single-stranded DNA and RNA molecules, such as single-stranded overhangs in otherwise double-stranded DNA or DNA-RNA molecules, or single-stranded loops in double-stranded molecules. Cleavage by S1 selectively removes single-stranded segments while leaving double-stranded segments intact. Although this method is suitable for the quantification of transcripts it is superior for mapping of 5′ and 3′ ends of transcripts and for the localization of introns.

Reaction Principle In the first step of the nuclease S1 analysis, a labeled, single-stranded DNA probe complementary to the RNA sought is hybridized to the isolated mixture of RNAs in solution. After adding nuclease S1, all single-stranded sections are digested and only the paired RNA-DNA hybrids remain. This double-stranded, cleavage-resistant RNA-DNA fragment is often referred to as the protected fragment. The labeled DNA fragments are then separated on an acrylamide gel and visualized by autoradiography or other suitable means, depending on the label employed. The resulting data provides two elements of information. The size of the remaining fragments allows the calculation of the distance to the end of the transcript or splice site. The intensity of the signal is proportional to the concentration of the complementary RNA species in the RNA mix and therefore provides information about the amount of the specific transcript.

Quantitative Nuclease S1 Analysis If the goal is to quantify an RNA species a complementary DNA probe can be chosen that hangs over the 5′ end of the RNA of interest. After hybridization, the nuclease S1 digests the overhang (Figure 34.1). DNA probes between 40 and 80 nucleotides long are particularly well suited for this assay. T4 polynucleotide kinase can be used to efficiently label the 5′ end of the DNA oligonucleotide. The specificity with which the hybridization probe is labeled is of decisive importance in determining the sensitivity of detection.

Alternatively, or in addition, a second labeled probe can be added to the sample in an equimolar amount to the first probe. This allows the direct comparison of the amount of two different RNA species in a single experiment, provided the resulting fragments are of differing length. To assure the validity of the RNA quantification, the hybridization probe must be always present in significant molar excess relative to the amount of RNA to be quantified and the hybridization must run to completion. The ideal hybridization conditions for a particular DNA probe and corresponding RNA samples must be determined empirically.

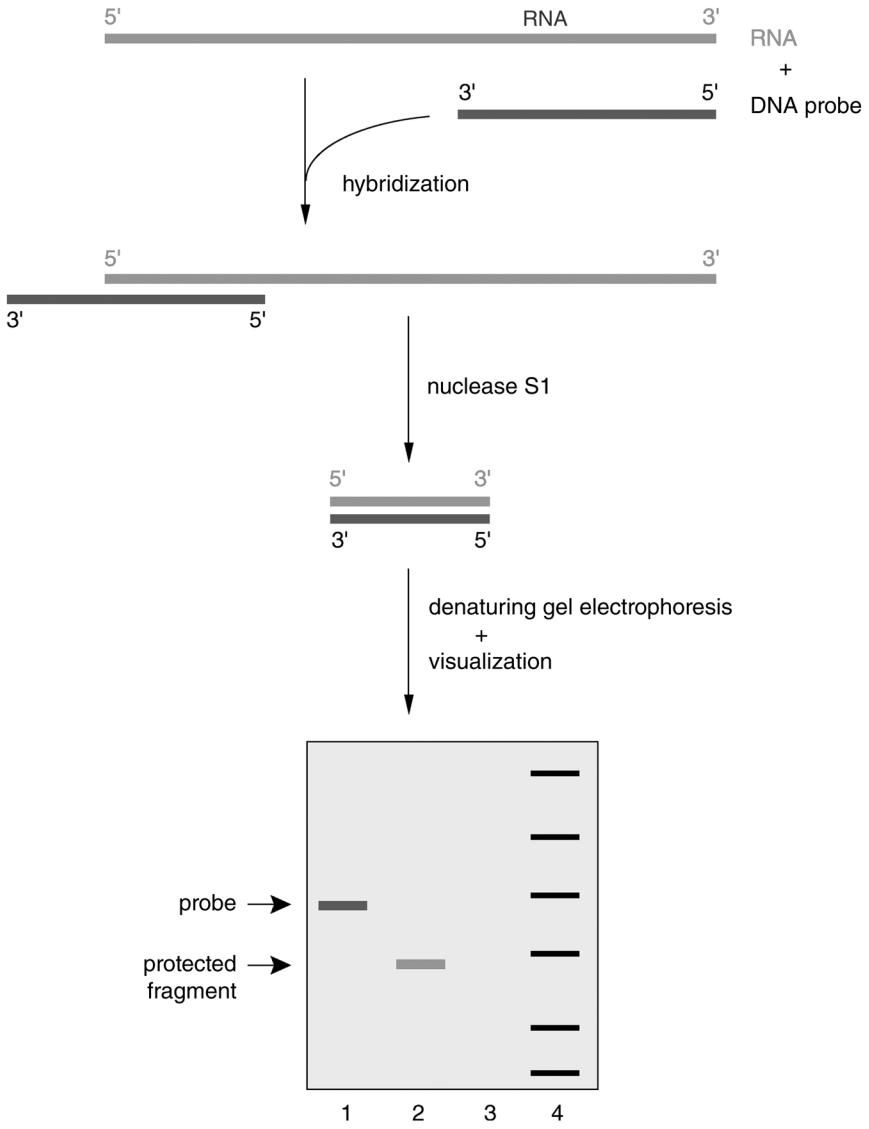


Figure 34.1 Principle of the nuclease S1 analysis for the quantification of RNA. The schematic diagram shows a typical result with the full-length probe in lane 1 and the protected fragment in lane 2. Lane 3 contains the negative control lacking RNA and lane 4 the nucleic acid size standards. A 5′-end labeled probe is used for this purpose.

The labeled DNA probe is present in significant excess in the reaction when, in the presence of a fixed amount of DNA probe, the signal resulting from a titration of differing amounts of RNA is proportional to the concentration of RNA. To determine if the hybridization reaction is complete, samples hybridized for varying length of time can be digested with nuclease S1. The reaction can be regarded as complete when the resulting signal no longer increases after longer hybridization periods.

Nuclease S1 Mapping of RNA 5′ and 3′ Ends Figures 34.2 and 34.3 show how nuclease S1 can be used to map the ends of an RNA and localize introns. Depending on the question a suitable hybridization probe must be selected. Should the 5′ end of a transcript be under investigation, a complementary DNA fragment is chosen that hangs over the 5′ end of the RNA and is labeled on its own 5′ end. After incubation with nuclease S1 the non-paired 3′ end of the DNA probe is degraded, resulting in a shortened fragment. When separated on a denaturing polyacrylamide gel, the difference in length between the original probe and the shortened fragment can be measured, which gives the distance between the 5′ end of the probe and the 5′ end of the RNA, the transcription start site. For this sort of high precision mapping a suitable size standard should be present on the same gel, ideally a Maxam–Gilbert DNA sequence reaction of the probe employed.

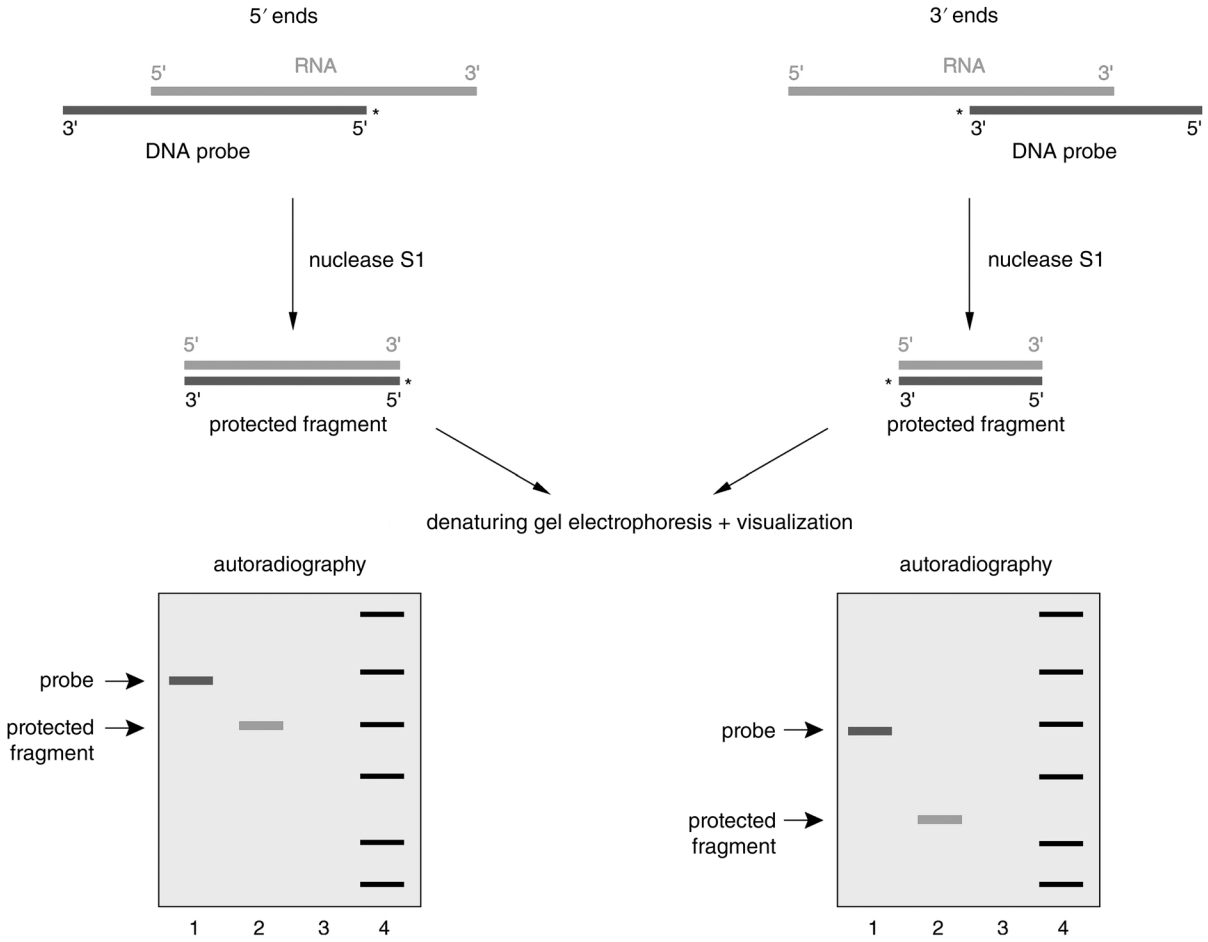


Figure 34.2 Mapping the 5′ and 3′ ends of an RNA. Lanes of the schematic are as described in Figure 34.1. Quantification makes use of a 5′ or 3′ end labeled probe.

A general problem of nuclease S1 mapping is that under sub-optimal reaction conditions nuclease S1 can hydrolyze double-stranded stretches of nucleic acids, which compromises the specificity of the method. For this reason it is important to always first optimize the reaction conditions in terms of reaction temperature and salt content.

PCR, Chapter 29

Mapping of the 3′ end of an RNA takes place analogously (Figure 34.2). The DNA probe used must be longer than the 3′ end of the corresponding RNA and the label must be on the 3′ end of the DNA.

To map exon/intron positions, a probe corresponding to the genomic sequence is labeled and hybridized to isolated RNA. If introns are present in the genomic sequence, the regions corresponding to exons pair with RNA but the regions corresponding to introns remain single-stranded loops (Figure 34.3). In the presence of nuclease S1 the single-stranded regions are digested, leaving the paired regions corresponding to the exons intact.

PCR can also be used to determine the intron/exon boundaries. PCR assays for the characterization of RNA species are carried out in parallel with sets of primer pairs on genomic DNA and reverse transcribed RNA. If the resulting products differ in length, this indicates the presence of exon/intron junctions in that segment.

34.1.3 Ribonuclease-Protection Assay (RPA)

The ribonuclease protection assay is an alternative to the nuclease S1 technique. Like the nuclease S1 analysis, RPA is based on the hybridization of isolated cellular RNA to a labeled nucleic acid probe, which is complementary to the RNA of interest. RPA is more sensitive than nuclease S1 analysis because the probe is labeled along its entire length and not just at its ends, and therefore has a higher specific activity. In addition, RNA-RNA hybrids are uncommonly thermostable.

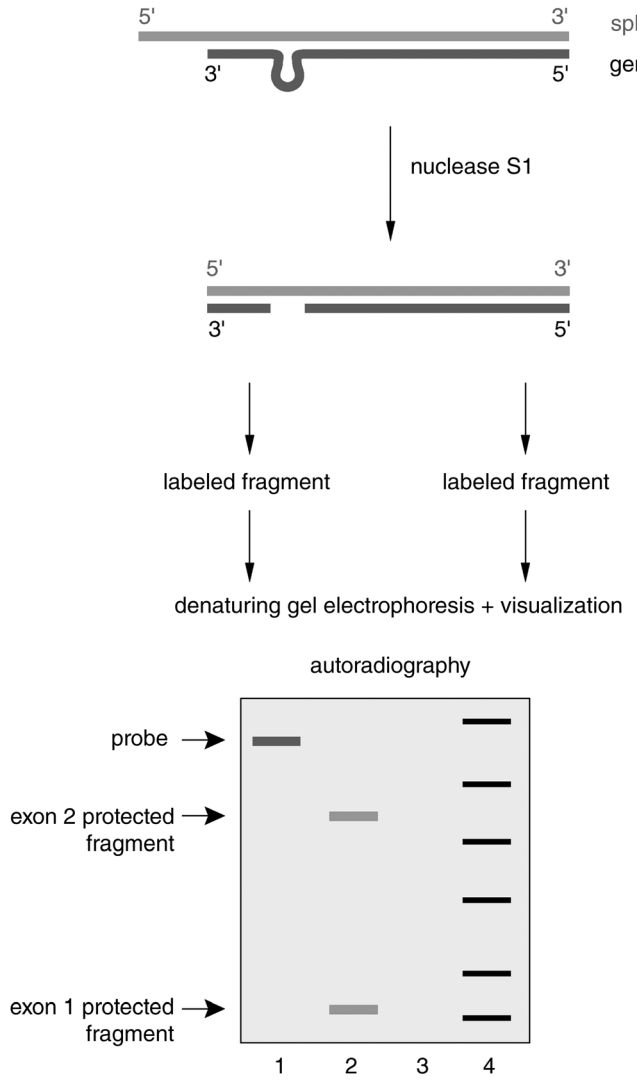


Figure 34.3 Locating an exon of a gene. Lanes of the schematic as described in Figure 34.1. Quantification makes use of a homogenously labeled probe.

As hybridization probe one needs an RNA probe complementary to the RNA of interest, generated by *in vitro* synthesis of an antisense probe, often referred to as the riboprobe. A prerequisite for the riboprobe is that the genomic area containing the gene segment of interest is cloned into a transcription vector. Transcription vectors typically contain specific promoters for one or more of three different bacteriophage RNA polymerases: SP6, T3, and T7. In such a vector it is easy to generate a run-off transcript of the region downstream of the promoter *in vitro* by adding the phage-specific RNA polymerase and the four ribonucleotides. This is illustrated in Figure 34.4. The resulting probe is useful not only for RPA, but can also be used as a hybridization probe in Northern blots.

RPA makes use of the labeled antisense RNA probe for the hybridization with isolated total RNA (Figure 34.5). The resulting perfectly paired RNA-RNA hybrid molecules are incubated with a mix of ribonucleases A and T1. Both nucleases specifically hydrolyze single-stranded RNA but differ in their specificity. RNase A, which is isolated from bovine pancreas, cleaves the phosphodiester bonds after pyrimidine nucleotides, RNase T1 from *Aspergillus oryzae* cleaves after guanine nucleotides. In the reaction, therefore, all single-stranded RNA regions are digested, including all the RNA molecules, which are not homologous to the RNA, single-stranded overhangs in the RNA-RNA hybrid, and the free, unbound probe. Only perfectly paired RNA-RNA hybrids consisting of cellular and probe RNA remain intact as protected fragments. These hybrids are subsequently separated

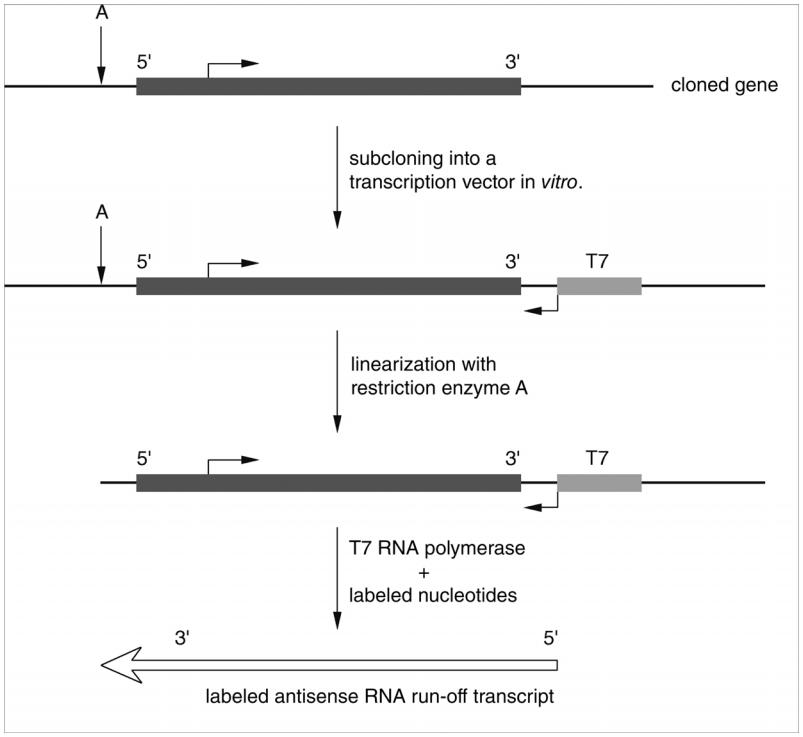


Figure 34.4 Creation of a riboprobe from a cloned gene.

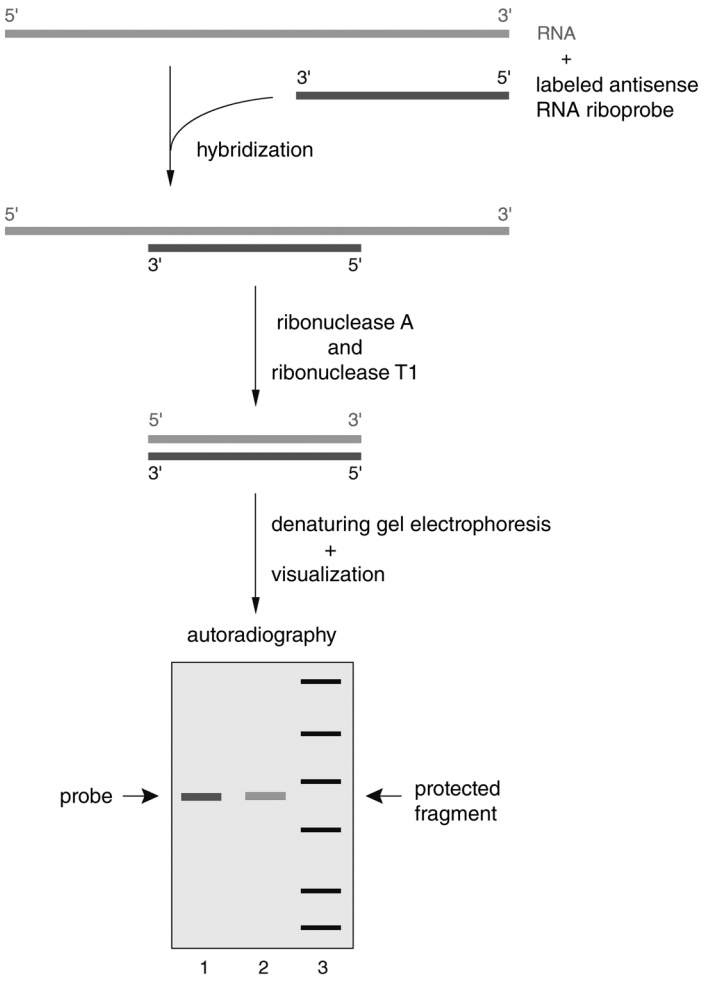


Figure 34.5 Principle of the ribonuclease protection assay. The schematic diagram shows the labeled probe in lane 1, the protected fragment in lane 2, and the nucleic acid size standards in lane 3.

on a gel, as for nuclease S1 assays, and visualized. Since the intensity of the resulting signal is proportional to the amount of hybridized cellular RNA, the amount of this RNA species in the reaction can be determined.

Besides quantification of a particular RNA species, this detection method can be used to map the end of transcripts. The same criteria discussed in Section 34.1.2 apply to the choice of hybridization probe except that a transcription vector must be used for the subcloning to allow generation of the probe.

RPA has several advantages relative to classical nuclease S1 analysis: The antisense RNA can be created in relatively large amounts and labeled with high specificity. RNA-RNA hybrids are significantly more thermostable than RNA-DNA hybrids, a key requirement for the creation of distinct protected fragments after nuclease treatment. Both play a significant role in increasing the sensitivity of this detection method: If end labeled probes are used in a nuclease S1 assay, using RPA will typically increase the sensitivity by between 20- and 50-fold. In addition, the hydrolysis with the ribonucleases is more reliable and reproducible, since, as previously mentioned, nuclease S1 reactions must first be optimized in terms of temperature, salt, and enzyme concentrations.

The high sensitivity of this method has, however, disadvantages. For example, if during the *in vitro* transcription reaction incomplete transcripts are generated in addition to the full length ones, later, after the RNase treatment, a mix of protected fragments will be observed. Incomplete transcripts can result when the bacteriophage RNA polymerase pauses on the template and the reaction terminates prematurely. Pausing is sequence dependent and can be avoided by ensuring that the DNA template chosen for the generation of the riboprobe is a relatively short gene segment between 100 and 300 bp in length.

34.1.4 Primer Extension Assay

Primer extension analysis is often used to determine the 5′ end of an RNA species and its amount in the same assay. The reaction is catalyzed by the RNA-dependent DNA polymerase reverse transcriptase (RT) in the presence of an RNA template and a DNA primer. The primer is single-stranded, complementary to a given RNA, has a length of about 20–40 nucleotides, and is radioactively labeled at the 5′ end using [^{32}P]-ATP and T4-polynucleotide kinase (PNK). In the first step of the reaction, this short ^{32}P-labeled DNA primer (5–10 pmol) is hybridized in solution with total or poly(A)-RNA (1–5 µg). After addition of RT and deoxynucleotides, the sequence of the DNA primer is elongated from its 3′-OH end by the enzyme, generating a cDNA (copy DNA), which is complementary to the RNA template bound by the primer. Ideally, the reaction stops at the 5′ end of the RNA template. The schematic in Figure 34.6 summarizes the steps of a primer extension assay. The newly synthesized 5′-end radiolabeled cDNA can be visualized following denaturing polyacrylamide gel electrophoresis (PAGE) by autoradiography or phosphorimaging.

The autoradiogram provides information about both the quantity and the length of the original RNA species. The amount of synthesized cDNA corresponds to the amount of the given RNA species in the initially isolated RNA pool. Therefore, a high copy number of the given RNA in the total RNA population leads to synthesis of a high number of cDNAs, which in turn results in a strong radioactive signal on the autoradiogram. On denaturing gels the length of the cDNA can be determined exactly with appropriately sized markers separated on the same gel. Thus, the information about the length of the cDNA and the position of the DNA primer reflects the distance from the primer to the 5′ end of the corresponding RNA. The position and specificity of primers used in this reaction must be optimized. The distance of the primer to the 5′ end of the corresponding RNA should not be more than 100 bases to avoid premature termination of the cDNA synthesis by the RT due to extensive secondary structures in the RNA. Multiple radioactive bands on the autoradiogram may indicate premature termination. However, in such a case, it is difficult to distinguish prematurely terminated cDNA from variable 5′ ends of the original RNA template. Indeed, more than one transcription start site of the corresponding gene will produce heterogeneous RNAs, which only differ in their 5′ end. In addition, RNAs encoded by multigene families may produce cDNAs of different lengths in a primer extension assay, if the target sequence selected for the primer is conserved among members of the gene family.

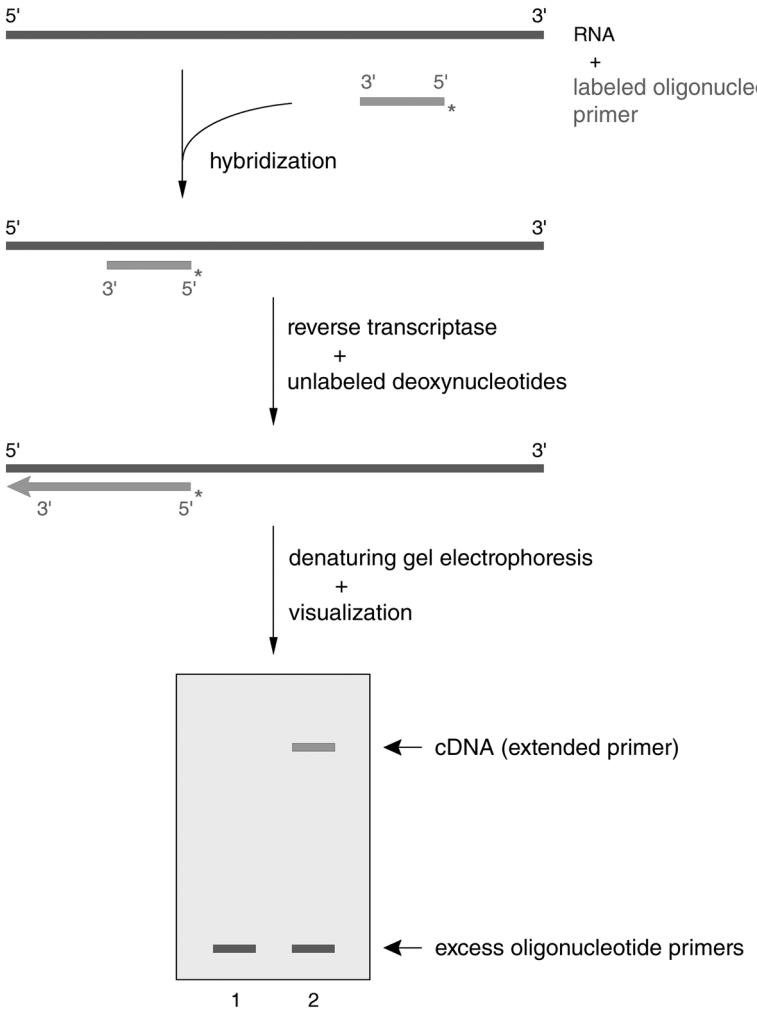


Figure 34.6 Principle of the primer extension assay. The schematic diagram shows the position of the end labeled oligonucleotide in lane 1, the cDNA primer extension product in lane 2, and excess labeled primer in both lanes.

Primer extension is frequently used for the detection of transcripts produced after transient transfection of cells or by *in vitro* transcription reactions in cell-free transcription systems. To detect variations in the amount of product not reflecting genuine differences between samples, it is important to plan controls into the experiments to detect variations in the efficiency of transfection, RNA isolation, and the efficiency of the primer extension reactions themselves. A heterologous RNA can serve as an internal control, which is detected with a second specific oligonucleotide probe in the same reaction, in addition to the primary probe used to detect the RNA of interest.

Mapping of 5′ ends of RNAs can be accomplished with two other methods previously described in this chapter. What are the basic differences? The nuclease S1 and ribonuclease protection assay are based on formation of stable RNA/DNA or RNA/RNA hybrids between the RNA and the complementary radioactive probe, respectively. Upon hybridization, non-complementary DNA or RNA is not hybridized and thus subject to digestion by the S1 nuclease (RNA/DNA) or by ribonuclease (RNA/RNA). The resulting protected fragments may represent either the true 5′ end of a processed RNA or an intron/exon boundary of the pre-mRNA. In contrast, primer extension by RT stops at the 5′ end of the RNA template after splicing has been completed, thereby delineating the 5′ end of the relevant RNA.

34.1.5 Northern Blot and Dot- and Slot-Blot

Northern Blot Northern blotting is a common and easy method to quantify a particular RNA within an RNA pool. This method is also described in Chapter 27, so only points related to RNA

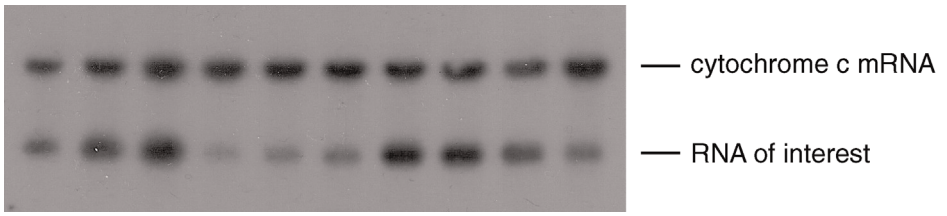


Figure 34.7 Example of RNA quantification by Northern blot. Each lane contains 10 µg total RNA. The figure shows an autoradiogram of the filter after RNA transfer and hybridization. The hybridization is carried out with a RNA probe complementary to the RNA of interest and a second probe complementary to cytochrome *c* mRNA to allow the results to be normalized.

quantification will be discussed here. For Northern blotting, either total cellular RNA or poly $(A)^+$-enriched RNA is denatured to unfold secondary structures in the RNA molecules, followed by electrophoresis on denaturing agarose gels to separate the RNAs according to their size. RNA is then transferred from the gels to nitrocellulose or nylon membranes. Specific RNAs are detected by membrane hybridization using a radioactive labeled RNA or DNA hybridization probe that is complementary to the relevant RNA.

Northern, Section 27.4.4

Northern blotting is useful to determine the relative amount, as well as the size, of a particular RNA, but not to define the precise 5′ or 3′ ends of the RNA. This technique is often used for analysis of a given RNA within a highly heterogeneous RNA pool, or RNA pools isolated from different sources (e.g., different cell lines or tissues). The comparison of the amount of a particular RNA from different RNA sources requires loading of equal amounts of total RNA as well as re-hybridization of the membrane with a normalization probe, specific to an RNA that is similarly expressed in many cell types and tissues to act as an internal standard. Usually, transcripts from so-called "housekeeping" genes – that is, genes that are believed to be expressed at a similar level in a wide range of cell types and tissues, irrespective of signaling pathways and cell differentiation – are used for normalization, such as cytochrome or actin mRNA. To avoid erroneous normalization, more than one housekeeping gene should be selected, since also expression of common housekeeping genes fluctuates in a signaling and cell differentiation dependent fashion. An example of RNA quantification by Northern blotting is given in Figure 34.7, showing the main principles of the analysis. The type and amount of RNA to be loaded depends on the abundance of the particular RNA species. Assuming an abundance of 0.1%, loading of 5–20 µg of total RNA is sufficient to allow quantification of the relevant RNA species. Within a total RNA preparation the amount of an mRNA of interest is pretty low, since the majority of RNA consists of rRNA (80%) and tRNA (15%). Therefore, an abundance of 0.1% of an mRNA of interest within a total RNA pool is very high. Consequently, for detection and quantification of low abundance mRNA species, that is, gene transcripts being present at low copy numbers (50 copies per cell), enrichment of poly$(A)^+$-RNA is required and up to 10 µg of poly$(A)^+$-RNA should be used.

In comparison to other methods, Northern blots require a large quantity of RNA for detection, are quantitatively inaccurate for low abundance RNAs, and are time consuming, which has led to the development of several techniques to get around these limitations.

Dot- and Slot-Blot Analysis Dot blot or slot blot are used to detect RNA in complex RNA samples, similarly to Northern blots, but without prior separation of the nucleic acids by electrophoresis. Instead, the RNA mixture is denatured with formamide, formaldehyde, and heat (65 °C) and then directly spotted onto nitrocellulose or nylon membranes under vacuum using a dot blot or slot blot apparatus. Fixation of the RNA on the membrane and detection of a particular RNA species using specific nucleotide probes is done as described for Northern blotting. This method is quick and often used to compare the amount of a specific RNA in many samples in parallel, for example, to determine expression patterns of genes in different tissues. Therefore, this technique is important for clinical diagnosis, such as monitoring expression of oncogenes. Real-time PCR has replaced this method for this purpose in many cases due to its greater speed and sensitivity.

Dot- and Slot-blot, Section 27.4.5

This technique gives no information about the size of the target RNA, since the initial RNA mixture is not fractionated by electrophoresis. To avoid errors due to cross hybridization of the radiolabeled hybridization probe to non-target RNAs, appropriate negative and positive control

samples should be analyzed in parallel. Moreover, quantification of the target RNA in the cellular RNA mixture requires quantification of known amounts of the target RNA synthesized *in vitro*.

34.1.6 Reverse Transcription Polymerase Chain Reaction (RT-PCR and RT-qPCR)

A powerful method for quantification of gene expression is transcription of total cellular RNA into cDNA by RT coupled with amplification of cDNA by PCR (RT-PCR) and gene-specific PCR primers. Currently, this technique is commonly used to monitor gene expression. In RT-PCR, total RNA is first converted into cDNA by RT, usually using random hexamers for priming cDNA synthesis instead of gene-specific primers. The cDNA pool is then used as template for amplification by PCR in the presence of gene-specific primers. RT-(q)PCR is the method of choice to measure even low gene expression accurately: Its high sensitivity allows detection of one specific RNA molecule among up to 10^8 unrelated RNA molecules. It has a wide range of clinical and non-clinical applications, including diagnosis of genetic diseases and cancer, analysis of gene expression patterns, or cloning of RNA. While RT-PCR is used for qualitative studies of gene expression or for cloning purposes, RT-qPCR (RT coupled with quantitative PCR) allows relative and absolute RNA quantification.

RT-PCR, Section 29.2.3

Determination of gene expression levels by RT-PCR techniques is only reliable if the amount of DNA amplified by PCR is proportional to the amount of RNA in the original sample material. Given that random hexamer-primed reverse transcription into cDNA should not be rate limiting, it is crucial to analyze the amount of PCR-amplified DNA during the exponential amplification phase, before the plateau phase of the PCR reaction is reached. In both methods, RT-PCR and RT-qPCR, quantification of a specific RNA is not possible without appropriate internal controls, in particular if expression patterns in samples from different biological materials are compared. To avoid errors in quantification due to variation in RNA preparation or cDNA synthesis efficiency between different samples, the amount of PCR products obtained for the RNA of interest is normalized to the amount of PCR product obtained for a reference RNA. Commonly, ribosomal RNA, actin mRNA, tubulin mRNA, or other transcripts from "housekeeping" genes are used for normalization. If accurate quantification is an absolute must, additional assays for the validation of results should be performed. Alternatively, an artificial RNA template is produced by an *in vitro* transcription assay. Such an artificial RNA contains the same primer binding sites as the natural RNA template but differs in length. This approach employs a cloned gene fragment encoding the RNA of interest, which is first mutated, either by a short deletion or insertion, then shuttled into an *in vitro* transcription vector just downstream of the promoter for RNA polymerases SP6, T3, or T7. Substitution of the *in vitro* transcription assay with the corresponding RNA polymerase (SP6, T3, T7) produces the artificial RNA. After purification and determination of the concentration of the *in vitro* synthesized RNA, it is reverse transcribed into cDNA along with the natural RNA, followed by PCR together with the natural RNA counterpart.

RT-PCR is usually performed as a two-step reaction. In the first step, RNA is incubated in assay buffer with non-specific primers, usually random hexamers, and RT. In the second step, PCR is performed with an aliquot of the cDNA reaction in the presence of gene-specific primers. Quantification of RT-derived cDNA can be achieved either by end-point PCR (RT-PCR) or by real-time RT-PCR (RT-qPCR). RT-qPCR is the method of choice to determine if gene expression changes globally and/or with many samples. Nevertheless, semi-quantitative end-point PCR is still used for measurement of gene expression in a small number of samples or if RT-qPCR is not available. Quantification of gene expression by relative RT-PCR requires co-amplification of an internal control together with the gene of interest to normalize the samples. As mentioned above, either synthetic RNAs or mRNAs from housekeeping genes are suitable as internal control. Upon normalization, relative abundances of transcripts can be compared across multiple samples. However, quantification is only reliable when both the target and control amplifications are within the linear range of the PCR reaction, which is usually determined in pilot experiments with serial dilutions of target and control cDNA or by trying different cycle numbers. Detection of the end-point RT-PCR products is done either by analysis on agarose gels and staining with ethidium bromide or by performing the PCR reaction in the

Quantitative PCR, Section 29.2.5

presence of ^{32}P followed by measurement of the radioactively labeled DNA by phosphor-imaging or by scintillation counting. The values are usually presented as ratios of the target gene signal to the internal control signal.

Real-time RT-PCR (RT-qPCR) has become the method of choice for quantification of gene expression, mostly replacing conventional RT-PCR and Northern blotting. The accumulation of PCR products is recorded during the individual amplification cycles, and not at the end of the reaction, by continuous measurement of fluorescent dyes that integrate or associate quantitatively with the amplified DNA. Two fluorescent dyes are widely used to quantify the accumulation of PCR products. SYBR-Green is a fluorescent dye that intercalates into any double-stranded DNA in a sequence-independent fashion. Alternatively, fluorescence reporter probes, such as *TaqMan* probes, are sequence-specific oligonucleotide probes that are labeled with a fluorescent reporter. During PCR, the probe hybridizes to the complementary sequence with the heterogeneous pool of PCR products, thereby quantifying individual amplicons. Since fluorometric detection of the labeled DNA is very sensitive, both approaches allow precise quantification of RNA (cDNA) in the original samples. Usually 10–100 cDNA copies per sample can be detected using SYBR-green or *TaqMan* probes in combination with the appropriate thermocycler model.

Hybridization Methods, Section 28.1.3

34.2 Analysis of RNA Synthesis *In Vivo*

All the previously described methods are suitable for measurement of the steady-state level of an RNA, but do not provide information about ongoing gene expression. Nuclear-run-on assays, however, allow the measurement of promoter activity and ongoing gene expression in living cells, thus facilitating studies on regulation of gene expression at the transcriptional level. This method was originally established to demonstrate that regulation of transcription initiation rates is the main rate-determining step in the synthesis of the non-mature precursor-RNAs in mammalian cells. Briefly, nuclear-run-on assays measure the synthesis of ongoing gene expression via incorporation of [^{32}P]-UTP into nascent RNA chains. Cells do not take up UTP efficiently from the culture medium. Therefore, cells are permeabilized under conditions that perforate the cell membrane, but do not destroy the nuclear membrane. Upon uptake, [^{32}P]-UTP is incorporated into initiated, nascent RNAs by the nuclear RNA polymerases I, II, and III. Since radiolabeling of the transcripts occurs during elongation, most of the transcriptome (steady state RNAs) is not labelled. As a result a nuclear-run-on assay measures the amount of ongoing transcription largely independent of RNA stability. This assay is important for addressing various scientific questions, such as regulation of gene expression during cell differentiation, at distinct phases of the cell cycle, or by chemical compounds.

Nuclear-run-on assays consist of the following steps:

- cell permeabilization/lysis,
- nuclear-run-on transcription (RNA labeling reaction),
- isolation of RNA,
- detection and quantification of the specific RNA.

Isolation of Genomic DNA, Section 26.2

Alternatively, a halogenated pyrimidine nucleoside, 5-fluoro-uridine (FU), bearing a fluoro substituent at position 5 on the uracil ring, is used for labeling of nascent RNAs *in vivo*. In contrast to [^{32}P]-UTP, FU is cell permeable, and permeabilization of the cell membrane is not required. FU-labeled transcripts are detected either by immunofluorescence microscopy or by RNA-immunoprecipitation (RNA-IP) using antibodies specific to FU.

34.2.1 Nuclear-run-on Assay

The following summarizes the important steps of nuclear-run-on assays using mammalian cells in culture. Cells are chilled on ice, and the cytoplasmic membrane is permeabilized or lysed. This treatment results in pausing of all RNA polymerases. Usually, cell lysis is performed in a hypotonic or isotonic buffer containing the non-ionic detergent NP-40 (NP-40 cell lysis buffer). The intact nuclei are pelleted by centrifugation, while components of the cytoplasm, as well as

Isolation of intact nuclei, which preserve their RNA polymerase activity, is essential for nuclear-run-on assays. Thus, cell lysis conditions and nuclei purification have to be optimized for a given cell type. Isolated nuclei are either used immediately for nuclear-run-on transcription or are snap-frozen in liquid nitrogen and stored at −80 °C in buffer containing glycerol.

Typically, a nuclear-run-on transcription contains between 5×10^6 and 5×10^7 nuclei per assay. Incorporation of [α-^{32}P]-UTP into nascent RNA drops significantly at low densities of the nuclei, which decreases the detection of gene-specific transcripts by reverse dot blotting. Under optimal assay conditions up to 10–30% of the input radioactivity will be incorporated into nascent transcripts during the nuclear-run-on transcription.

the cytoplasmic membrane, remain in the soluble supernatant. Depending on the cell type used, efficient cell lysis may require homogenization with a Dounce homogenizer. In addition, further purification of nuclei is achieved by centrifuging nuclei through a sucrose cushion. Nuclei are then incubated for a short time (pulse labeled) at 30 or 37 °C in the presence of radiolabeled [α-^{32}P]-UTP and non-labeled ribonucleoside triphosphates (NTPs). During this step, new transcripts are not initiated, but [α-^{32}P]-UTP will be incorporated into virtually all nascent transcripts that were initiated at the time point when the cells were chilled and lysed. Finally, DNA and proteins are digested with RNase-free DNase and proteinase K, and RNA is extracted with guanidinium thiocyanate–phenol–chloroform extraction (TRizol).

Quantification of the relative amount of nascent transcripts in each sample involves a modified dot (or slot) blot technique (Section 34.1.5). In this *reverse* dot blot procedure, a non-labelled DNA probe containing the gene of interest is immobilized on a nitrocellulose filter or nylon membrane and hybridized to the purified [^{32}P]-labeled RNA synthesized during the nuclear-run-on transcription. The membrane-bound DNA probe should be in high molar excess to prevent saturation of the DNA probe during the hybridization reaction, otherwise the calculations underestimate the amount of RNA synthesized. The amount of radioactivity that hybridizes to the membrane is approximately proportional to the number of nascent transcripts. Since the number of nascent transcripts on a gene is thought to be proportional to the frequency of transcription initiation, the amount of hybridized ^{32}P-labeled RNAs reflects the activity of the corresponding gene promoter.

34.2.2 Labeling of Nascent RNA with 5-Fluoro-uridine (FUrd)

Genomic loci with high transcriptional activity can be visualized by short-term labeling of nascent RNA with the non-radioactive uridine analogues 5-fluoro-uridine (FU) or BrUTP (BrU) (Figure 34.8). Both analogues can substitute for UTP during transcription of nascent transcripts and are not incorporated into mature RNA. Labeled RNA is subsequently detected by immunochemical methods. In contrast to BrUTP, fluoro-uridine is efficiently taken up by intact cells, and no permeabilization of cells is necessary. Upon uptake, fluoro-uridine is quickly metabolized to fluoro-UTP, the NTP required for incorporation into RNA by the nuclear RNA polymerases I, II, and III. BrUTP is not cell permeable and so additional procedures are required to facilitate uptake, such as permeabilization of the cytoplasmic membrane or purification of nuclei (as described above), microinjection, or use of liposome particles. Fluoro-uridine based pulse labeling of nascent RNA is gentle and quick since the entire cell remains intact.

For *in situ* labeling, FU (2 mM) is directly added to the cell culture medium for 5–30 min. After labeling, cells are fixed, permeabilized, and prepared for indirect immunofluorescence microscopy. Detection of nascent FU-labeled RNA is accomplished by incubation with a primary antibody specific to fluoro-uridine or bromo-uridine, followed by incubation with an appropriate secondary fluorescently labeled antibody. Fluorescence signals are finally visualized under a microscope. If labeling is done for a very short time, visualization of nascent transcripts is restricted to genomic loci with the highest amounts of RNA polymerase activity. Under these conditions, usually only transcripts in nucleoli are detectable. Nucleoli represent the nuclear body, where rDNA is transcribed into rRNA from arrays of more than 150 repeated gene copies by highly active RNA polymerase I (Figure 34.9). Extension of the labeling time up

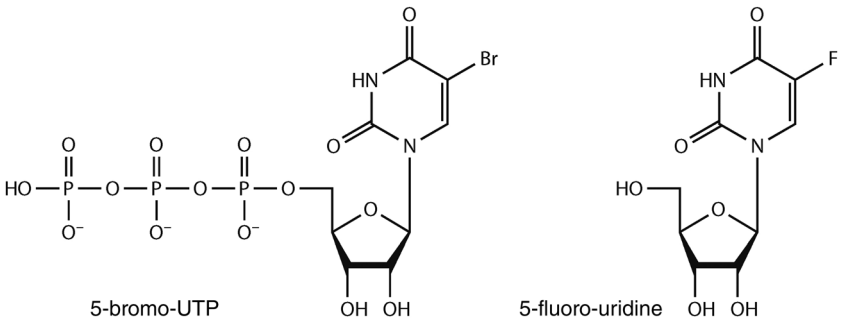


Figure 34.8 Structures of 5-bromo-UTP (BrU) and 5-fluorouridine (FU).

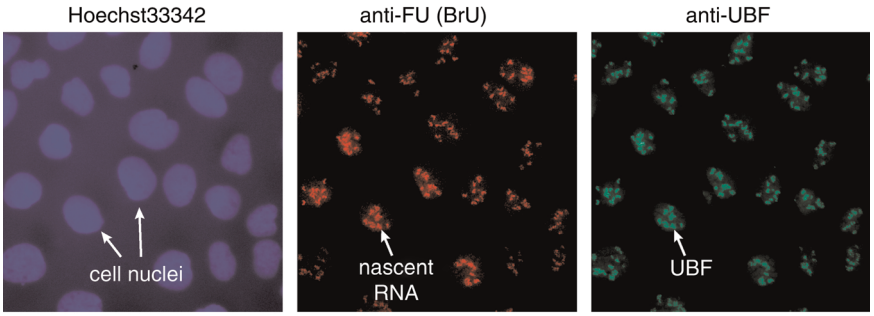


Figure 34.9 *In vivo* labeling of nascent RNA by 5-fluoro-uridine. Human U2OS cells were cultured in the presence of 2 mM 5-fluoro-uridine (FU) for 10 min. Cells were then fixed and permeabilized. FU-labeled RNA was detected by immunofluorescence microscopy using a monoclonal antibody against FU/BrdU and a secondary Cy3-labeled anti-mouse antibody (red signals). The visualized signals correspond to nascent nucleolar RNA, since they co-localize with the nucleolar transcription factor UBF visualized by rabbit anti-UBF and secondary FITC-labeled anti rabbit antibodies (green signals). The DNA is counterstained with the dye Hoechst 33342 (blue signal).

to 30 min allows visualization of FU-labeled mRNAs as well, which becomes evident through the increasing accumulation of fluorescence signals in the nucleoplasm.

Alternatively, antibodies specific to FU can be used to immunoprecipitate FU-labeled nascent transcripts. Immunoprecipitated RNA is extracted by guanidinium thiocyanate–phenol–chloroform, and purified RNA is analyzed by either RT-qPCR (Section 34.1.6) or RNA-seq.

34.3 *In Vitro* Transcription in Cell-Free Extracts

Transcription initiation is controlled by a multitude of factors, including general transcription factors, gene specific factors, and accessory co-regulators. The interplay between sequence-specific binding of transcription factors to their binding sites in proximity of the core promoter and the sequential assembly of general transcription factor complexes around the core promoter and transcription start site (TSS) is necessary for gene-specific recruitment of RNA polymerase I, II, or III and accurate transcription initiation. The level of transcription initiation is highly variable, depending on the level and activity of gene-specific transcription factors. Often, the amount and activity differ according to cell type, cell cycle, differentiation state of a cell, and intra- and extracellular signaling pathways. Moreover, *in vivo,* transcription initiation depends on the accessibility of transcription factors to the gene regulatory cis-acting sequence elements, which are embedded in chromatin. Thus, changing the chromatin structure by chromatin remodelers and histone modifying complexes is crucial for regulation of transcription. Many questions related to regulation of gene expression can be addressed by *in vitro* transcription of cloned gene promoters in cell-free extracts or reconstituted transcription systems. This experimental approach helps to identify protein components that activate or repress transcription, as well as relevant cis-acting regulatory elements on the corresponding gene of interest.

34.3.1 Components of an *In Vitro* Transcription Assay

In vitro transcription needs an appropriate DNA template, which typically contains the core promoter, gene control elements located upstream of the core promoter, and a stretch of the transcribed region located downstream of the transcription start site (TSS). These *cis*-acting sequence elements are cloned into a plasmid, resulting in the final plasmid construct, which is called a minigene reporter. Since the plasmid is amplified and purified from *Escherichia coli*, it is nucleosome-free. Transcription of the minigene *in vitro* requires a protein extract that contains the minimal set of protein components essential for transcription, the ribonucleotides ATP, CTP, GTP, UTP, and radiolabeled [α-^{32}P]-UTP, a defined set of cations (Mg^{2+}, K^{+}, Zn^{2+}, and others, depending on the gene context), and buffer substances (Tris, HEPES). Given that all components of the assay system are active, exact initiation and elongation of transcription *in vitro* can be assessed.

To decipher regulatory cis-elements in the gene of interest, deletions, insertions, or point mutations can be introduced into the cloned gene fragment by i*n vitro* mutagenesis. *In vitro* transcription of mutated promoter constructs is performed in transcription-competent extracts, and the amount of reporter transcript is compared to the level of transcript synthesized from the non-mutated wild-type minigene. Similarly, up- and down-regulation of transcription by gene-specific transcription factors can be assessed. For this purpose, the transcription factor of interest should be depleted from the protein extract, to determine if depletion up- or down-

regulates *in vitro* transcription. After the depleted protein extract is reconstituted with the depleted protein or protein complex, transcription should be restored to levels observed in the non-depleted extract if the depleted protein/protein complex was the only factor removed from the protein extract. Such a transcription assay also facilitates identification of functional domains in transcription factors, if transcriptional activity is compared in extracts reconstituted with the wild-type factor or with a mutant protein.

34.3.2 Generation of Transcription-Competent Cell Extracts and Protein Fractions

Extracts for *in vitro* transcription assays are commonly prepared from cultured cells. Transcription extracts are divided into three categories depending on the method used for cell lysis and fractionation: cytoplasmic extracts, nuclear extracts, and whole cell extracts. Usually, extracts prepared from isolated nuclei can be used directly for functional studies, including *in vitro* transcription and RNA processing. Moreover, nuclear extracts are the starting material for purification of proteins involved in these processes. Depending on the cell type used for extract preparation, individual steps have to be optimized to generate reproducibly active transcriptional extracts and to increase the yield of components necessary for transcription. Briefly, cells are incubated in hypotonic buffer, the swollen cells are lysed by Dounce homogenization, and the intact nuclei are collected by centrifugation. Nuclei are then resuspended in high salt buffer to extract soluble proteins from the nuclei. All buffers are free of detergents to prevent inactivation of components of the transcription machinery. All components of the nuclei that are resistant to high salt extraction (e.g., chromatin) are pelleted by centrifugation, while nuclear transcription factors and RNA polymerases remain in the supernatant (nuclear extract). After dialysis of the nuclear extract into a moderate salt solution it is ready to use for *in vitro* transcription. Alternatively, the nuclear extract can be used as starting material for biochemical enrichment of factors by ion exchange-, gel filtration-, or affinity-chromatography.

34.3.3 Template DNA and Detection of *In Vitro* Transcripts

Accurate transcription *in vitro* depends on active cell extracts or a well-defined set of proteins fractionated by biochemical methods and a DNA template that contains all essential promoter elements of the corresponding gene. If an important element is missing, synthesis of *in vitro* transcripts will be impaired, regardless of whether active extracts or protein fractions were used in the assay system. Two types of construct are widely used: (i) DNA templates that contain a G-less cassette (Figure 34.10) and (ii) DNA templates that are linearized by restriction enzymes downstream of the transcribed region producing *in vitro* run-off transcripts (Figure 34.11).

The G-less Cassette The G-less cassette is often used to determine the promoter strength of a gene of interest *in vitro*. To generate a G-less cassette DNA template, the promoter, together with upstream regulatory elements, or the core promoter alone is inserted into a plasmid containing the G-less cassette. Usually, a G-less cassette consists of a synthetic DNA fragment of 350 bp lacking guanine residues in the sense direction. The promoter is cloned immediately upstream of the G-less cassette. The G-less cassette is transcribed in the *in vitro* system using the constituents described above (Section 34.3.1) except that GTP is replaced by the GTP analogue 3′-*O*-methyl-GTP. Upon transcription initiation, RNA polymerases transcribe the G-less cassette until the first guanine residue. At the first guanine downstream of the G-less cassette, 3′-*O*-methyl-GTP is incorporated into the nascent transcript, which prevents further elongation of the nascent RNA chain and causes termination of the transcript. The 350 nt labeled transcript is released from the template DNA (Figure 34.10) and is subjected to polyacrylamide gel electrophoresis. The amount of label incorporated into the transcript is quantified by autoradiography, phosphorimaging, or other suitable means, and indicates the relative strength of the promoter of the gene of interest. Transcriptions with G-less cassette templates can be performed on circular plasmids. In many transcription systems circular plasmids are more efficiently transcribed than linearized templates, which are used for *run-off* transcription. In addition, the presence of 3′-*O*-methyl-GTP suppresses promoter-independent random and non-specific transcription throughout the plasmid.

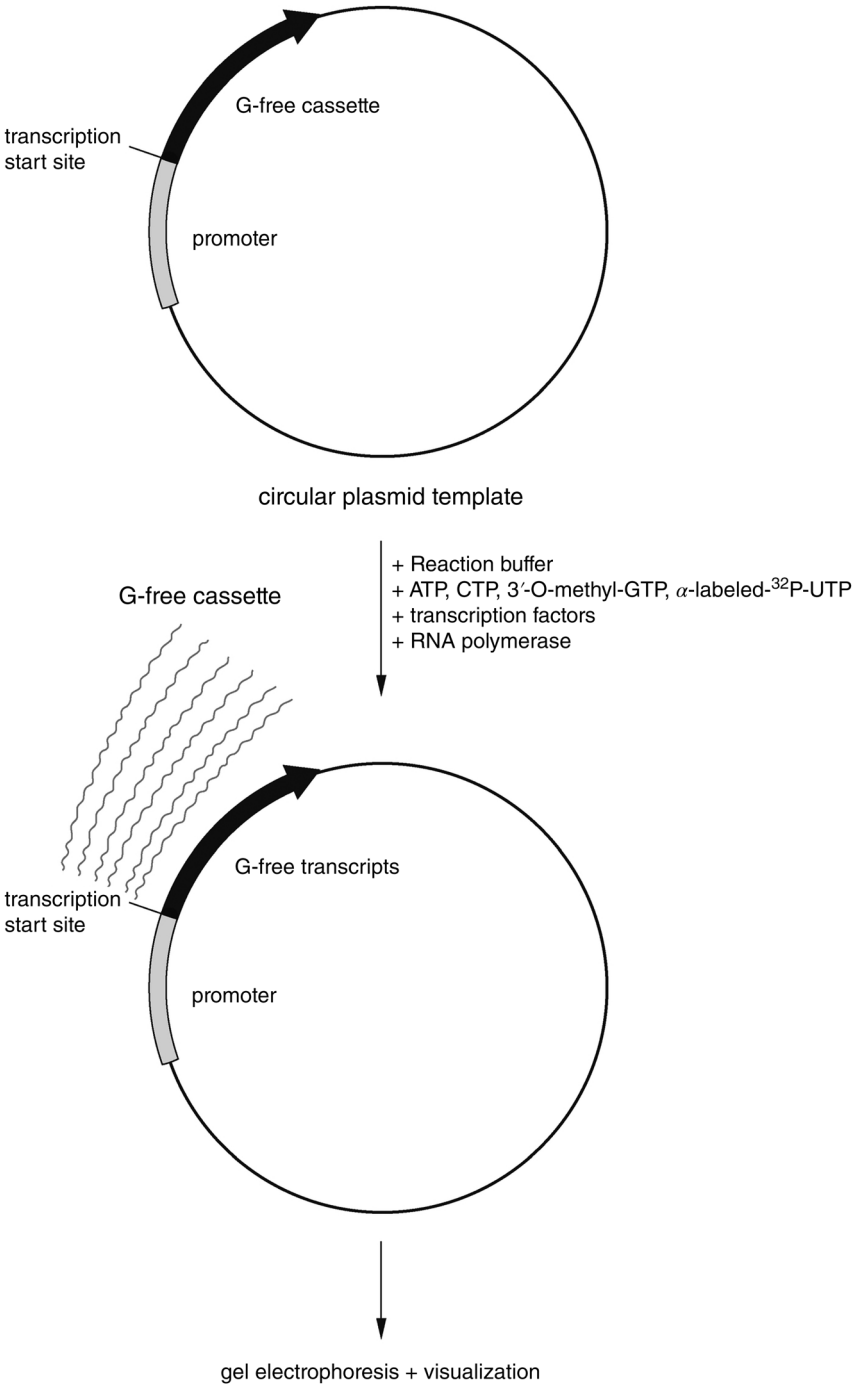


Figure 34.10 Use of G-less cassettes for *in vitro* transcription. The promoter-containing plasmid is incubated with extract or partially purified transcription factors and RNA polymerase, ribonucleotides, and reaction buffer. GTP is replaced by 3′-*O*-methyl-GTP in the reaction, leading to chain termination once incorporated into the RNA. The resulting run-off transcripts are subsequently separated on a gel and detected.

Run-off Transcripts For use in *run-off* transcription assays, the promoter of the gene of interest and regions downstream of the transcription start site (TSS) are cloned into a plasmid. Termination of transcription is achieved by linearization of the plasmid at a defined restriction enzyme cleavage site. Upon initiation, transcripts are synthesized until the RNA polymerase arrives at the 3′ end and "runs off" the linearized template. Since the distance between TSS and the restriction enzyme cleavage site is defined, a ^{32}P-labeled *run-off* transcript of the predicted size indicates specific initiation of transcription. The size of the RNA is verified by gel electrophoresis. As before, the intensity of the signal indicates the relative amount of RNA produced and reflects the strength of the promoter (Figures 34.11 and 34.12).

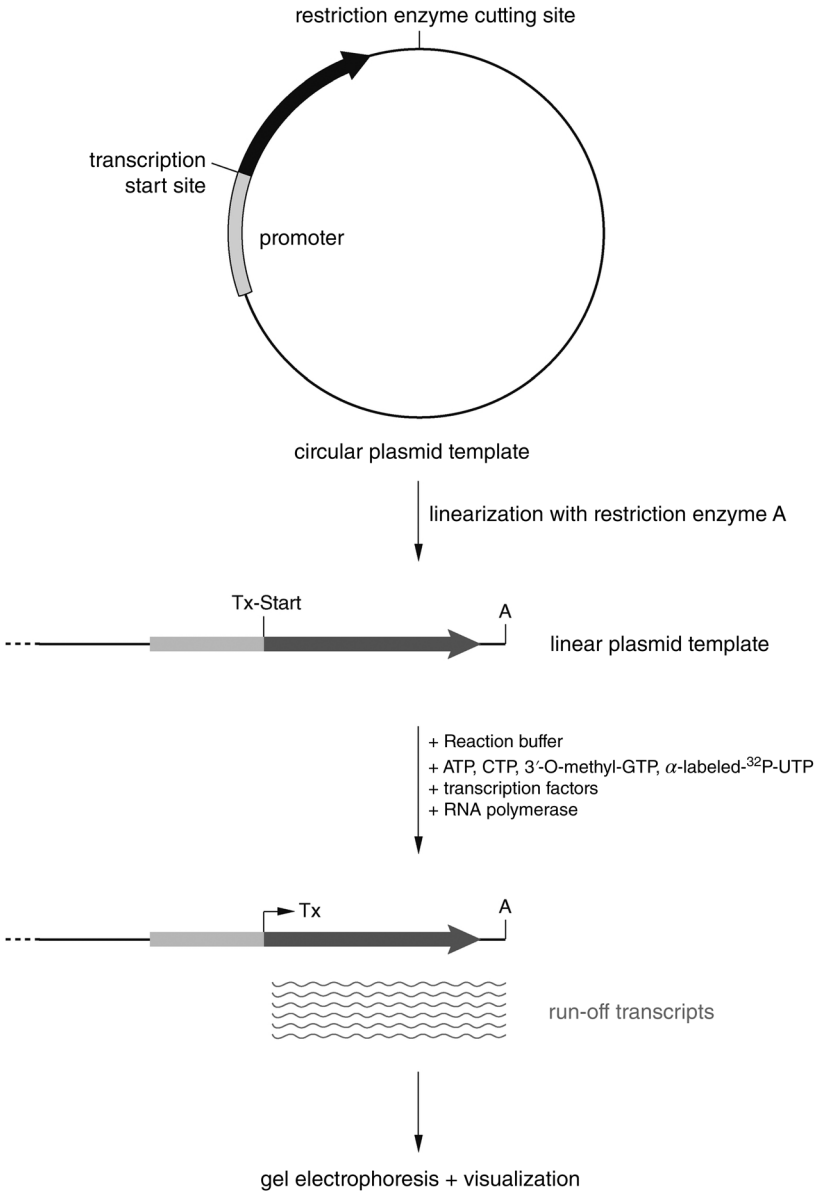


Figure 34.11 Principle of the *in vitro* run-off transcription reaction. The promoter containing plasmid is linearized with a restriction enzyme that has a unique recognition site and is located at an appropriate distance (usually 150–500 bp) downstream of the transcription start site. The linearized template is incubated with extract or partially purified transcription factors and RNA polymerase, ribonucleotides, and reaction buffer, and the resulting run-off transcripts are analyzed as given in Figure 34.10. Tx, transcription.

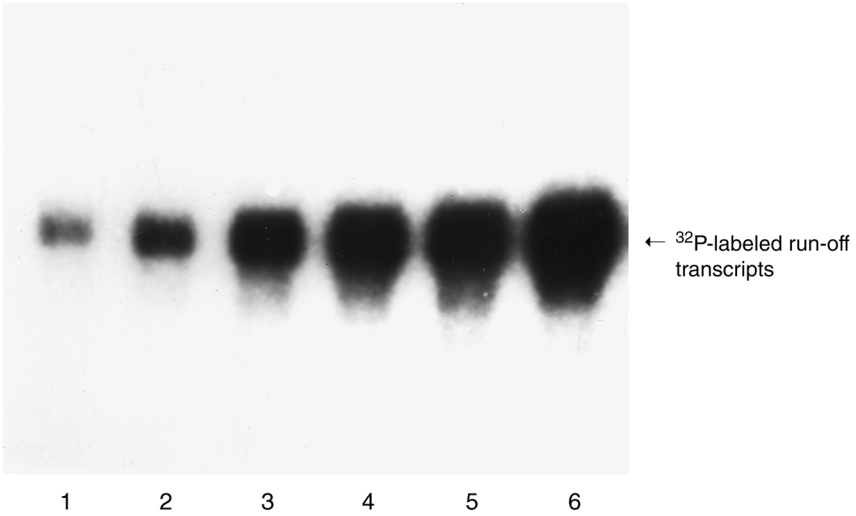


Figure 34.12 Example of an *in vitro* run-off transcription analysis. A linearized template containing the rDNA promoter was incubated with increasing amounts (lanes 1–6) of essential transcription factors and RNA polymerase I.

Additional Techniques to Analyze *In Vitro* Transcripts Sometimes, it is not feasible to use G-less cassettes as DNA template for *in vitro* transcription. In addition, direct detection of ^{32}P-labeled run-off transcripts on polyacrylamide gels is only possible if strong promoters are tested, for example, rDNA promoter and promoters for viral genes. If a promoter is weak, the levels of spurious transcripts may be as high as promoter-independent transcripts. In such cases, *in vitro* transcripts should be analyzed by more laborious techniques described above (Section 34.1) to qualitatively and quantitatively validate the RNA: nuclease S1 mapping, ribonuclease S1 protection, primer extension assays.

34.4 *In Vivo* Analysis of Promoter Activity in Mammalian Cells

Many questions relating to promoter activity cannot be definitively solved by *in vitro* transcription approaches (e.g., changes of promoter activity by growth factors and signaling molecules) but should be assessed *in vivo* in their proper cellular context. In addition, the significance of gene-regulatory cis-acting sequence elements identified *in vitro* has to be validated *in vivo*. Analysis of promoter activity in mammalian cells consists of three main steps: (i) cloning of the regulatory elements of a gene of interest into an appropriate transfer or reporter gene plasmid, (ii) introduction of the cloned plasmid into mammalian cells, and (iii) assays aimed at quantifying the activity of the cloned promoter.

34.4.1 Vectors for Analysis of Gene-Regulatory cis-Elements

Identification of promoter elements and promoter-proximal regions that up- or down-regulate the corresponding gene commonly require cloning of promoter elements into an appropriate plasmid vector, which is introduced into mammalian cells, followed by measurement of promoter strength. Usually, vectors employed for this purpose are reporter genes.

Isolation of Plasmid DNA from Bacteria, Figure 26.6, Section 26.3.1

The vector backbone of such a plasmid has several functional parts (Figure 34.13):

- bacterial origin of replication (for propagation in *E. coli*),
- antibiotic resistance gene (selection marker),
- multiple cloning site (MCS),
- reporter gene sequence cloned into the MCS,
- polyadenylation signal sequence downstream of the cloned reporter sequence;
- optional elements are:
  - second antibiotic resistance for selection in mammalian cells,
  - weak minimal promoter immediately upstream of the reporter gene,
  - eukaryotic origin of replication (for propagation in mammalian cells).

Reporter genes are used to assess the activity of promoters and additional gene regulatory DNA elements of a gene of interest (target) in cultured mammalian cells. For this purpose elements of the target gene promoter are cloned immediately upstream of the reporter gene coding sequence. After introduction into cells, the reporter gene is expressed under the control of the target gene promoter. The amount of reporter gene product is measured and the results are normalized to the amount or activity of the reporter gene produced by a reference promoter (Figure 34.13). Commonly used reporter genes are luciferase (Luc), β-galactosidase (β-Gal), green fluorescent protein (GFP), and chloramphenicol acetyltransferase (CAT). Detection methods used to measure the expressed reporter gene product involve luminescence, fluorescence, thin-layer chromatography (TLC), or RNA analysis (Section 34.4.3).

Mutational Analysis of Promoters Promoter malfunction has been associated with hundreds of diseases, malfunction being often caused by mutation of a promoter sequence or upstream regulatory elements. To decipher the underlying molecular mechanisms it is important to test individual promoter elements of a candidate gene, either alone or in combination, in reporter gene

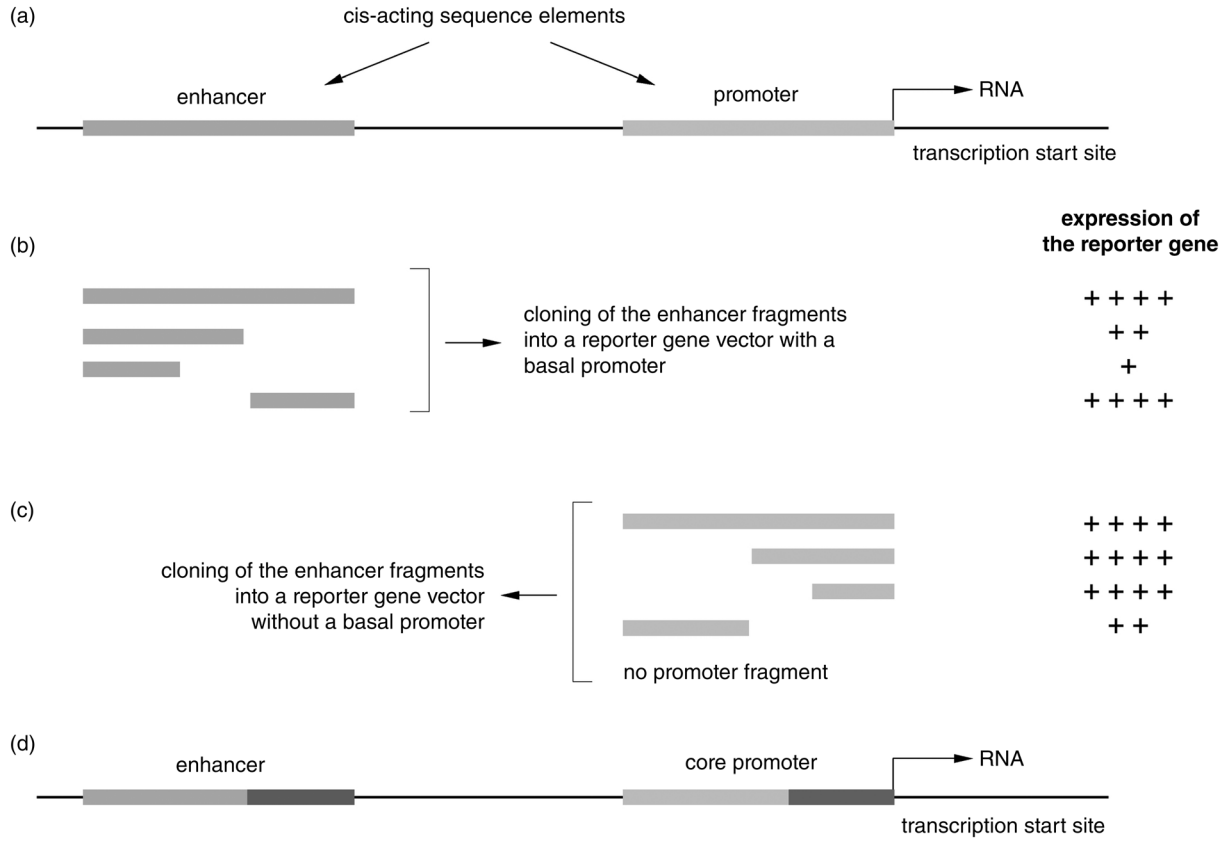


Figure 34.13 Principle of mapping the important cis-acting sequences with the aid of a reporter gene vector. (a) Promoter structure of the hypothetical gene. Enhancer, proximal promoter, and the transcription start site are shown. (b) After restriction analysis or sequencing, the entire enhancer region or fragments thereof are cloned into the multiple cloning site of an eukaryotic expression vector containing a basal promoter and reporter and, after transfection, the reporter gene activity of the cell lysates is measured (++++ = high; + = low). (c) To analyze the minimal proximal promoter, the same basic procedure is followed as described for (b). Here, however, a promoter-less plasmid with the reporter gene is used. Without a functioning promoter, no reporter protein is made (–); if important regulatory sequences are missing, the rate of reporter gene synthesis is low (++). (d) From the results of the experiments described in (b) and (c), the functionally important enhancer regions (dark gray) and promoter regions (red) can be elucidated.

assays for gain-of-function or loss-of-function. Based on mutational analysis, functionally important nucleotides of sequence elements can be identified (Figure 34.13). Commonly, nucleotide point mutations are introduced by site-directed mutagenesis, and sequence stretches are deleted or introduced by PCR and/or restriction enzyme digestion and re-cloning of the modified DNA fragment into the reporter gene plasmid. A gene "enhancer" has no intrinsic promoter activity but activates (enhances) expression of the corresponding gene in cis, though sometimes at a considerable distance from the promoter and orientation-independent. To test enhancer activities, putative enhancer elements are cloned into a reporter gene plasmid that contains a eukaryotic reference promoter. Viral promoters (SV40, RSV, CMV) are usually used for this purpose. Active enhancer elements stimulate reference promoter activity by increasing expression of the reporter gene product.

34.4.2 Transfer of DNA into Mammalian Cells

Foreign DNA can be introduced into mammalian cells in two ways, either by transfection or with the help of viruses as vehicles (viral transduction). For each category, several techniques have been established. Transfection of mammalian cells is based on a transient increase of the permeability of the cytoplasmic membrane, which allows uptake of the foreign nucleic acids or modified gene plasmids by the cell. Non-viral introduction of DNA is achieved by chemical-based transfection, non-chemical-based techniques, or lipofection. Notably, the efficiency of transfection is highly variable between different cell types and the proper method is best determined empirically on a case-by-case basis.

Transient and Stable Expression Upon uptake, the vast majority of transfected plasmids are not integrated into chromosomal DNA, which results in the transient expression of the cloned genes (transient transfection). The plasmids are diluted in each round of cell division or degraded. However, transfection of plasmids containing viral origins of replication, such as the Epstein–Barr virus (EBV) or SV40, allow episomal amplification of the plasmids in appropriate daughter cells: The EBV origin of replication requires cells expressing the EBV encoded nuclear antigen 1 (EBNA1) and the SV40 origin of replication requires cells expressing the SV40 large-T-antigen. Routinely, HEK293E or HEK293T cells are used for this purpose. Episomal amplification greatly reduces dilution of the plasmid during cell division.

In very few cells (approximately 1 out of 10^4) the foreign plasmids will have randomly integrated into the chromosomal DNA of the host cell (stable transfection). To accomplish selection and enrichment of the few stably transfected cells, the plasmid backbone of eukaryotic expression plasmids usually encodes a drug resistance gene. Since this drug resistance gene is co-integrated into the chromosomal DNA, the drug resistance gene is constitutively expressed. If cells are cultured in the presence of the drug, only those few cells will survive that express the resistance gene, whereas other cells will die. Common drugs used for selection of stable transfected cells are geneticin (G418), puromycin, or zeocin.

Chemical-Based Transfection Chemical-based transfection is widely used for the introduction of nucleic acids into a broad range of mammalian cell types. One of the most popular techniques for transfection of cells growing in a monolayer is based on *calcium phosphate*. The target DNA is mixed with a solution of HEPES-buffered saline (HBS) containing phosphate and calcium chloride. When all three components are combined, a precipitate of positively charged calcium ions, negatively charged phosphate ions, and the DNA is formed. Upon addition to cells in culture, cells take up the DNA–calcium phosphate crystals by a process that is not completely understood. Depending on the cell type, transfection efficiencies of up to 90% or more have been observed. Alternatively, cationic polymers are used such as *DEAE-dextran* or polyethyleneimine (PEI). The cationic polymers form a complex with the negatively charged DNA, which cells take up by endocytosis.

Non-chemical Methods

Electroporation is a popular and efficient technique to transfect mammalian cells growing in suspension or adherent cells that are resistant to transfection by chemical methods. In this technique, a suspension of cells and the DNA to be transfected are exposed to short pulses of intense electricity using a special electroporation device. Electric pulses of several hundred volts are applied that transiently increase the permeability of the cell membrane, allowing introduction of the DNA into the cells. Expression plasmids can be mechanically introduced into cells or nuclei by microinjection with the use of a glass micropipette under a microscope. This technique is only convenient for analyses of a limited number of cells, although automated systems are available to improve the handling.

Lipofection A broad range of mammalian cells is efficiently transfected by means of liposomes. Lipofection uses cationic liposomes, often in combination with neutral co-lipids, as a carrier for the DNA. Positively charged liposomes and negatively charged DNA form complexes, with the positive charge on the surface and the packaged DNA inside. These DNA–liposome complexes fuse with the cell membrane, releasing the packaged DNA into the cell.

Viral Transduction DNA can also be introduced into mammalian cells using viruses as carrier (viral transduction). Delivery of a gene of interest by a virus requires cloning of the corresponding gene into a viral vector. Transduction efficiency is extremely high compared to alternative transfection methods, often close to 100%. In addition, transduced genes are often integrated into the chromosomal DNA of the infected cell, facilitating stable gene expression. Viral vectors have been engineered for use in basic research and gene therapy. Widely used are recombinant retroviruses, lentiviruses (a subclass of retroviruses), adenoviruses, and adeno-associated viruses. The following key features are common to all viral vectors, (i) low toxicity, which reduces undesired side effects on the physiology of the infected cell; (ii) high stability of the viral genome after integration into the genome of the infected cell, preventing frequent gene rearrangements; (iii) cell type specificity, to ensure infection of a broad range of cell; and (iv) marker genes, for identification of cells infected by

the virus (i.e., antibiotic resistance genes). In particular, viral transduction is applied for delivering recombinant genes into cells, in which transfection with the aforementioned techniques is inefficient, such as mouse embryonal fibroblasts (MEFs). Widely used are retroviruses due to the property that the recombinant retroviral vector integrates into the mammalian host genome at high frequency. Upon integration, the recombinant gene is stably propagated to daughter cells of the host. The gene of interest is first cloned into a retroviral vector. Such vectors are genetically modified but retain all gene elements, which are required for replication of the recombinant retroviral vector in the mammalian host cell as well as integration into the mammalian genome. Safety is of high priority to minimize risk of handling the viral vectors. Usually, parts of the viral genome necessary for production of infectious viral particles are deleted from the viral vector. Therefore, production of infectious viral particles requires that the recombinant viral vector is transfected into a packaging cell line that encodes the missing proteins for assembly of infectious recombinant viruses, such as envelope proteins. During this process, a recombinant viral stock is generated that can be used to infect appropriate host cells and introduce the gene of interest.

34.4.3 Analysis of Reporter Gene Expression

Reporter genes assays have become an indispensable tool in studying gene expression. They are widely used in basic, biomedical, and pharmaceutical research. Reporter gene proteins are expressed under the control of the promoter and/or promoter response elements of interest. Following expression, the cells are assayed for the reporter protein, either by direct determination of its amount or by measurement of its enzymatic activity. Ideal reporter gene proteins are easily detectable and their activity directly correlates with their expression level. This ensures that expression of reporter gene protein reflects the strength of the promoter under investigation. Some reporter gene systems can be used to measure transcriptional activity of the promoter of interest i*n vivo*, that is, if a promoter is fused to the green fluorescent protein (GFP). The basic principles of widely used reporter genes are described below.

Chloramphenicol Acetyltransferase (CAT) Assay In the CAT reporter system, the reporter gene encodes the enzyme chloramphenicol acetyltransferase (CAT). CAT is a bacterial enzyme that detoxifies the antibiotic chloramphenicol by acetylation and confers chloramphenicol resistance to bacteria. Acetyl-CoA dependent acetylation of chloramphenicol (Cm) by CAT produces 1,3-diacetylated chloramphenicol and mono-acetylated 3-acetyl and 1-acetyl intermediates, which are all biologically inactive. The activity of CAT in cell lysates is monitored by acetylation of radioactive labeled [^{14}C]-chloramphenicol. Non-acetylated and acetylated forms of [^{14}C]-chloramphenicol are separated by thin-layer-chromatography (TLC). The ratio between the acetylated derivatives, which are more mobile, and non-acetylated chloramphenicol reflects the CAT activity and the level of gene expression (Figure 34.14).

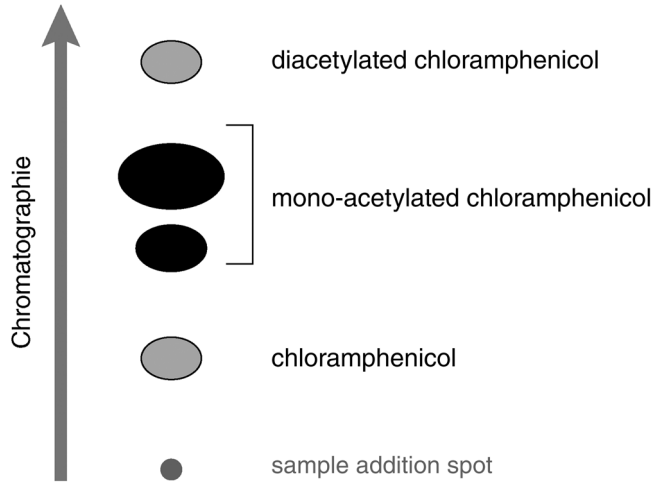


Figure 34.14 CAT assay: schematic diagram of the thin-layer chromatographic separation of the reaction products resulting from a CAT assay.

Luciferase Assay The luciferase gene from the firefly *Photinus pyralis* is the most popular reporter gene. Luciferase is an enzyme that oxidizes D-luciferin in the presence of ATP, oxygen, and Mg^{2+}, yielding CO_2 and a fluorescent product (oxyluciferin) that can be quantified by measuring the amount of emitted light. Photon emission is detected using a light sensitive luminometer. Since light excitation is not required for luciferase bioluminescence, background fluorescence is extremely low. Therefore, luciferase-based reporter assays allow accurate quantification of subtle changes in gene expression due to high sensitivity and a broad linear range of the enzymatic reaction. The activity of less than 0.1 pg of luciferase can still be accurately measured in a standard reaction.

Non Radioactive Systems: Bioluminescence, Section 28.4.3

β-Galactosidase Assays β-Galactosidase (β-Gal) is an enzyme that catalyzes the hydrolysis of β-galactosides into monosaccharides, and, in addition, hydrolyses several non-physiological substrates. Non-physiological substrates are used to measure β-galactosidase activity in cell extracts. Depending on the substrate, β-galactosidase activity is determined by colorimetric or fluorometric assays, or by chemiluminescence. In the basic colorimetric assay, *O*-nitrophenyl-β-D-galactopyranoside (ONPG) is used as an artificial chromogenic substrate. ONPG is colorless, while the product of the reaction, *o*-nitrophenol (ONP) is yellow ($\lambda_{max} = 420$ nm). Therefore, β-galactosidase activity is measured by the rate of appearance of the yellow color using a spectrophotometer. Although ONPG is the most commonly used substrate, the sensitivity of the colorimetric assay is low. Fluorescence-based assays, which utilize substrates that fluoresce upon hydrolysis, provide increased sensitivity. The substrate 4-methylumbelliferyl-β-D-galactopyranoside (4-MUG) does not fluoresce until cleaved by β-galactosidase, generating the fluorophore 4-methylumbelliferone (4-MU). The production of the fluorophore is monitored at an emission/excitation wavelength of 365/460 nm. Most sensitive is the chemiluminescence assay provided by 1,2-dioxetane substrates (i.e., 1,2-dioxetane-galactopyranoside derivatives). β-Galactosidase catalyzes decomposition to 1,2-dioxetane, which emits light with a maximum intensity at a wavelength of 475 nm. Light production is measured with a luminometer. The chemiluminescence-based assay allows detection of less than 1 pg of β-galactosidase in the reaction.

Although ß-galactosidase expression can be used as a standard reporter for monitoring the strength of a promoter or enhancer, it is now predominantly used as an internal control during transfection experiments. When used in this manner, cells are usually transfected with the control plasmid expressing β-galactosidase under the control of a viral promoter, such as the SV40 promoter, together with a second plasmid containing another reporter gene (e.g., luciferase or chloramphenicol acetyltransferase) under the control of the promoter or enhancer of interest.

Green Fluorescent Protein Expression of green fluorescent protein (GFP) is a widespread tool used to visualize spatial and temporal gene expression patterns *in vivo*. Details regarding green fluorescent proteins are described in Chapter 7. Usually, GFP gene reporters are not employed to quantify gene expression levels.

Green Fluorescent Protein (GFP) as a Unique Fluorescent Probe, Section 7.3.4

Analysis of the Transcripts from Transfected Cells In addition to, or rather than, reporter gene protein analysis, the amount of mRNA transcribed specifically from the test promoter can be quantified. Many of the methods presented above in Section 34.1 are suitable for this purpose, in particular quantitative RT-PCR, but also nuclease S1 assays, ribonuclease protection assays, and Northern blots are useful. Reporter mRNAs differ from native RNAs in that they are usually not subject to post-transcriptional regulation and are usually considerably more stable than native RNAs. This has the desirable effect of amplifying signals, but in so doing also only shows part of the full picture affecting the expression and regulation of an mRNA. For a native mRNA subject to regulation by miRNAs, for example, measuring the reporter will completely ignore the effects of this regulation. This is useful, in that it allows the researcher to separate promoter activity from mRNA regulation, but if the question relates to regulation of mRNA levels it provides only a partial understanding. Current research clearly indicates that post-transcriptional regulation is more frequent than originally thought, though also limited by the fact that it rarely affects levels by more than about 2–4-fold and never acts as a clear on–off switch, like a promoter can.

Further Reading

Brown, T., Mackey, K., and Du, T. (2004) Analysis of RNA by northern and slot blot hybridization. *Curr. Protocols Mol. Biol.*, unit 4. 9. Doi: 10.1002/0471142727.mb0409s67.

Carey, .F., Peterson, C.L., and Smale, S.T. (2013) The RNase protection assay. *Cold Spring Harbor Protocol*, issue 3.

Chen, J.L. and Tjian, R. (1996) Reconstitution of TATA-binding protein-associated factor/TATA-binding protein complexes for *in vitro* transcription. *Methods Enzymol.*, **273**, 208–217.

Cornetta, K., Pollok, K.E., and Miller, A.D. (2008) Retroviral vector production by transient transfection. *Cold Spring Harbor Protocols*, issue 4.

Cornetta, K., Pollok, K.E., and Miller, A.D. (2008) Generation of stable vector-producing cells for retroviral vectors. *Cold Spring Harbor Protocols*, issue 4.

Eyler, E. (2013) Explanatory chapter: nuclease protection assays. *Methods Enzymol.*, **530**, 89–97.

Johnson, G., Nour, A.A., Nolan, T., Huggett, J., and Bustin, S. (2014) Minimum information necessary for quantitative real-time PCR experiments. *Methods Mol Biol.*, **1160**, 5–17.

Fu, Y. and Xiao, W. (2006) Study of transcriptional regulation using a reporter gene assay. *Methods Mol. Biol.*, **313**, 257–264.

Ko, H.Y., Hwang, do W., Lee, D. S., and Kim, S. (2009) A reporter gene imaging system for monitoring microRNA biogenesis. *Nat Protocol*, **4**, 1663–1669.

Li, L. and Chaikof, E.L. (2002) Quantitative nuclear run-off transcription assay. *Biotechniques*, **33**, 1016–1017.

McPheeters, D.S. and Wise, J.A. (2013) Measurement of in vivo RNA synthesis rates. *Methods Enzymol.*, **530**, 117–135.

Nolan, T., Hands, R. E., and Bustin, S.A. (2006) Quantification of mRNA by real-time RT-PCR. *Nat. Protocol*, **1**, 1559–1582.

Percipalle, P. and Louvet, E. (2012) In vivo run-on assays to monitor nascent precursor RNA transcripts. *Methods Mol. Biol.*, **809**, 519–533.

Romero-Lopez, C., Barroso-del Jesus, A., Mendendez, P., and Berzal-Herranz, A. (2012) Analysis of mRNA abundance and stability by ribonuclease protection assay. *Methods Mol. Biol.*, **809**, 491–503.

Sambrook, J. and Russell, D.W. (2006) Mapping RNA with nuclease S1. Cold Spring Harbor Protocols, issue 1.

Sambrook, J. and Russel, D.W. (2006) Calcium-phosphate-mediated transfection of eukaryotic cells with plasmid DNAs. *Cold Spring Harbor Protocols*, issue 1.

Sambrook, J. and Russel, D.W. (2006) DNA transfection mediated by lipofection. *Cold Spring Harbor Protocols*, issue 1.

Sambrook, J. and Russel, D.W. (2006) DNA transfection by electroporation. *Cold Spring Harbor Protocols*, issue 1.

Schmittgen, T.D. and Livak, K.J. (2008) Analyzing real-time PCR data by comparative C(T) method. *Nat. Protocol*, **3**, 1101–1108.

Schnapp, A. and Grummt, I. (1996) Purification, assay, and properties of RNA polymerase I and class I-specific transcription in mouse. *Methods Enzymol.*, **273**, 233–248.

Smale, S.T. (2010) Luciferase assay. *Cold Spring Harbor Protocols*, issue 5.

Smale, S.T. (2009) Nuclear run-on assay. *Cold Spring Harbor Protocols*, issue 11.

Southern, M.M., Brown, P.E., and Hall, A. (2006) Luciferase as reporter genes. *Methods Mol Biol.*, **323**, 293–305.

Spector, D.L. and Goldman, R.D. (2006) Constructing and expressing GFP fusion proteins. *Cold Spring Harbor Protocols*, issue 1.

Tuschl, T. (2006) Transfection of mammalian cells with siRNA duplexes. *Cold Spring Harbor Protocols*, issue 1.

Venkatesh, L.K., Fasina, O., and Pintel, D.J. (2012) RNase mapping and quantification of RNA isoforms. *Methods Mol Biol.*, **883**, 121–129.

Fluorescent *In Situ* Hybridization in Molecular Cytogenetics

35

Michelle Neßling and Karsten Richter

Service Unit Electron Microscopy, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

Molecular cytogenetics aims to characterize the genomic state of clonally dependent cell populations, for example, to determine aberrations that lead to tumorigenic devolution, to disclose gene-defects of unborn children by prenatal diagnosis, or to reveal the level of genomic relationship among species. Like classic karyotyping, fluorescent *in situ* hybridization is able to reveal numeric (ploidy, polysomy) and structural (translocation) chromosomal aberrations. However, the potential of the method extends further to high-throughput capability for determination of loss or gain of chromosomal regions. The resolution is also significantly better, though diagnosis of point mutations requires other techniques.

Physical and Genetic Mapping of Genomes, Chapter 36

To reveal aberrations in the genome of a cell, classic cytogenetics resort to the structural analysis of banding patterns from karyograms (karyotyping). Serious drawbacks of this approach are: (i) Only dividing cells are addressable by karyotyping. (ii) Identification of aberrations from the ideotype requires excellent expertise. (iii) Resolution is poor: single bands represent more than 1 Mbp of a genome, while more than one band is required to identify a translocated piece of chromatin. Molecular cytogenetics solves these problems: as pioneering work by John *et al*. and Pardue & Gall in 1969 demonstrated, genomic sites of interest can be shown by their hybridization with a labeled marker-DNA (probe).

Karyogram – Ordered representation of all chromosomes from a single cell nucleus, prepared as so-called metaphase-spread and labeled according to Giemsa, which results in a chromosome-specific banding pattern. Both numeric and structural aberrations relative to the standard banding pattern of the organism (Ideogram) are discernable to the level of a few bands.

Here, two modifications used to diagnose genomic aberrations by fluorescent hybridization will be discussed, namely, *in situ* hybridization (ISH) and comparative genomic hybridization (CGH). ISH uses fluorescence-labeled oligonucleotides as reporter for their target-sequence within a sample, for example, to show gene-loci within interphase nuclei or chromosome regions in a metaphase-spread. In contrast, CGH reverses this principle: the genome to be investigated is prepared and fluorescently labeled, and hybridized to an immobilized target of known spatial composition.

35.1 Methods of Fluorescent DNA Hybridization

35.1.1 Labeling Strategy

Fluorescence as a signal (FISH: Fluorescent-ISH) is now established against older strategies using radioactivity or enzymatic color-reactions. The fluorescent dark field signal is very sensitive, it is quantitative, and a large gamut of fluorochromes is available to label different targets in parallel with different colors. Ratio labeling, in contrast, demands exhaustive exposure times and yet responds with inferior spatial resolution. Furthermore, security issues demand for the management of an expensive hot-laboratory. DNA probes linked to enzymes, instead of to fluorochromes, had been employed to reveal hybridization targets by a downstream enzymatic reaction that produces colored precipitations (e.g., using horse-reddish peroxidase to produce brown diaminobenzidine). Such color precipitates are observed in

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

bright field, showing the histology of the sample in usual counterstain. However, colored-precipitations by enzymatic reaction are far less sensitive than fluorescence, and quantitative evaluation is approximate only.

Detection Systems, Section 28.4

Two types for fluorescent labeling are commonly distinguished, *direct* and *indirect labeling,* according to the strategy used to couple the reporter (fluorochrome) to the probe. Direct labeling with fluorochromes covalently bonded to the probe is particularly beneficial for multicolor applications to reveal multiple targets in parallel. Instead of a fluorochrome, a hapten may be bonded covalently to the probe. A second reaction then is required to reveal the hapten by a fluorescent reporter. Two commonly used systems are biotin/fluorescent streptavidin and digoxigenin/fluorescent anti-digoxigenin. An important advantage of indirect labeling is the potential to enforce signal strength via cascades of second marker reactions.

Technical parameters that influence the signal strength are the length of the oligonucleotide probe, the density of reporter-nucleotides per probe, the efficiency of hybridization along the target sequence, and the suppression of background through nonspecifically bound reporter.

35.1.2 DNA Probes

Technical demands for hybridization increase on going from repetitive to singular probes. Repetitive probes target clusters of repetitive genomic sequences (e.g., centromeric satellite DNA). Sequences addressed by these probes can be very short (below 1Mb) and still yield strong signals. False positive background is not an issue. In contrast, singular probes, which are designed to label extended genomic regions, provoke a strong false positive background, which needs to be suppressed to obtain meaningful signals. Since unique probes need to cover quite long stretches of target sequence to produce measurable signals they typically include interspersed repetitive sequences (IRSs), which exist throughout the entire genome, such as short and long interspersed nuclear elements. Section 35.1.4 gives details of a special approach called CISS that is used to suppress genomic IRS for hybridization with singular probes. This kind of background suppression is particularly important for the application of painting probes used to study the spatial extension of particular chromosome bands or arms or even whole chromosomes throughout a cell's genome.

Isolation of Genomic DNA, Section 26.2

Many probes used for medical applications are commercially available. Furthermore, probes may be tailored by applying dedicated primer-extension strategies to genomic DNA. Modern synthetic primer design profits from the quick development of online data bases as indispensable tools for the search of adequate sequences (NCBI, Ensembl, and UCSC). In former times, genomic libraries had been managed to provide specified DNA-fragments (ImaGenes, EMBL, Sanger Center, NIH), amplified in various vector-systems, based on phages, cosmids, BACs (bacterial artificial chromosomes), and PACs (P1 bacteriophage artificial chromosomes). Painting probes for whole chromosomes and chromosome arms and bands, for example, were collected from DNA libraries. Such libraries may be initialized from flow-sorted chromosomes or even needle-scratches from chromosomes on conventional metaphase-spreads, which then need equal amplification (e.g., using DOP-PCR).

The haploid genome of human beings weighs about 3 pg at a cumulated length of 3×10^{12} bp.

CGH-probes comprise the entire DNA of a cell-population under study. Even small tumor-biopsies yield enough DNA for an experiment. DNA from single cells and small cell populations, however, needs to be amplified. Since CGH measures the balance between two DNA-pools (e.g., tumor versus wild-type), amplification strategies must have an equal efficiency over the whole genome.

35.1.3 Labeling of DNA Probes

Methods of Labeling, Section 28.3

Hybridization probes get their reporters upon incubation with modified nucleotides during synthesis via for example, nick-translation, random priming, or the polymerase chain reaction (PCR). The probe length should range between 300 and 800 bp. Long probes are sterically hindered to efficiently pervade the target, while short probes tend to hybridize less specifically. Both cases will increase the background of labeling. One special advantage of nick translation is the potential to control the probe length by the efficiency of the DNase-I digest. After synthesis, free unbound reporter nucleotides must be cleared off the probe-sample (e.g., by ethanolic

precipitation of the probe DNA or by column-filtration). One hybridization experiment consumes about 10–100 ng of probe DNA.

Nick Translation Nick translation is based upon polymerase-I (Pol-I) dependent incorporation of nucleotides at single-strand breaks (nick). The first reaction step is a DNase-I digestion of the double stranded DNA to obtain single-strand breaks at an appropriate rate. Second, starting at each nick, polymerase-I will exchange one half-strand in the 5′-direction, deleting the old stand by its 5′-exonuclease activity while polymerizing new nucleotides to the 3′-OH residue at the nick, thus allowing incorporation of reporter nucleotides. In this way, the nick moves in the 5′-direction until another nick on the complementary strand of the DNA is encountered, causing a double-strand break. The length of labeled DNA-fragment therefore depends on the frequency of nicks and can experimentally be controlled by the intensity of DNase-I digestion.

Random Priming This is the method of choice for labeling fragments of initial DNA that are too short for nick-translation (i.e., below 2 kb). This DNA pool becomes denatured into single strands to be used as templates for primer-extension by Klenow Pol-I polymerization. The required primers are offered as a mixture of hexanucleotides in all possible combinations of the four bases.

PCR Labeling This is the natural way for labeling probes that are obtained by PCR amplification.

35.1.4 *In Situ* Hybridization

Strategies towards high-specificity hybridization for FISH and CGH proceed in four distinct steps: denaturation (probes and target-DNA), pre-annealing (saturation of ubiquitous repetitive sequences), hybridization, and stringency differentiation.

Further required measures are the improvement of target-accessibility (e.g., acid histone-extraction, protein digestion, and detergence extraction), as well as additional labeling steps, for example, involved with indirect labeling and for counterstaining chromosomes or nuclei.

Specificity of the Hybrization and Stringency, Section 28.1.2

Denaturation The denaturation of DNA occurs a few degrees above its melting temperature Tm. Simple boiling for some minutes will do to melt probe-DNA. However, if structural integrity is of concern, too much heat is not acceptable. This is typically applicable to denature target-DNA. To diminish Tm, incubation-buffers include high concentration of monovalent salt plus 50–70% formamide as a destabilizer.

Pre-annealing This is a strategy used to suppress background from IRS (interspersed repetitive sequences), which is necessary for the detection of unique target sequences by singular probes. After melting the labeled probe-DNA, non-labeled $C_0t - 1$ DNA is added in excess and allowed to "pre-anneal" with complementary sequences of the probe DNA, which lose their potential to hybridize to corresponding sites of the target (chromosomal *in situ* suppression, CISS).

C*$_0$*t is a value used to characterize the complexity of a given DNA pool and is derived from its typical duration for re-hybridization. The part of a metazoan genome with $C_0t < 1$ represents its repetitive sequences.

Hybridization The hybridization of probe DNA to the target DNA is initiated on lowering the temperature below T_m. To guarantee saturation, conditions are set to low stringency. A reaction time of a few hours is allowed for the hybridization of simple repetitive probes while CGH experiments with complex singular probes require several days.

Stringency Another source of background is the hybridization of partners that match only occasionally. The probability for these kinds of partnerships is intrinsically high with painting experiments. Two strategies are useful to suppress this effect: (i) adding non-labeled DNA of another species (e.g., salmon sperm DNA) to the hybridization-mix and (ii) appending a stringency wash following hybridization. Poorly paired partners become re-dissociated in washing buffers of diluted salt concentration and at increased temperature.

35.1.5 Evaluation of Fluorescent Hybridization Signals

Light Microscopy Techniques – Imaging, Chapter 8

DNA Microarray Technology, Chapter 37

FISH and CGH experiments, which are performed for their complex structural information, are investigated by fluorescence microscopy. Microarray CGH, in contrast, by design allows full automatic data-acquisition by a simple chip-reader. In any case, fluorescence is measured – the choice of fluorescent markers evidently must account for the technical capabilities of the instrumentation available (filter settings and laser types). In this respect, multicolor applications are particularly demanding. For such cases, it is worth determining whether or not physical overlap of signals occurs. Thus, to perform CGH, signals are identified unequivocally by their spectral signature (multicolor FISH, Section 35.2.1), as chromosome paints do not overlap in metaphase spreads. The use of the same probes to evaluate vicinal relationships of chromosomes in interphase, in contrast, requires sufficient spectral separation of the detection channels (Figure 35.1) to omit false signaling due to cross-talk. Though, in principle, cross-talk can be eliminated analytically, this will inevitably reduce the signal-to-noise ratio.

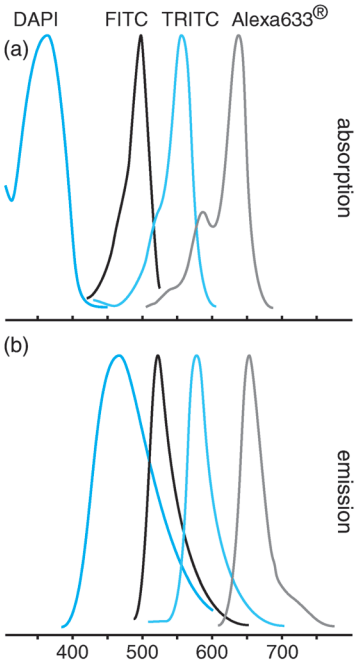


Figure 35.1 Comparing the absorption (a) and emission-spectra (b) of DAPI, FITC, TRITC, and Alexa633®. All four spectra are separated sufficiently to allow unequivocal allocation of four non-overlapping targets. However, overlapping is also notable: TRITC is significantly excited at 488 nm, the excitation-maximum of FITC, and emission of DAPI still occurs in the detection range of FITC.

35.2 Application: FISH and CGH

The aim of FISH is to localize a target within its native environment: The target sequence is known and the experiment designed to localize its abundance, for example, to map a genomic region throughout the genome of a tumor. As a consequence, FISH refers to single-cell events and many observations are to be collected for representative appraisal of a cell population.

CGH, in contrast, is designed to reveal numeric genomic aberrations of a cell-collective (e.g., tumor sample). A known target (metaphase spread of a standard cell line, sequence-collection on a DNA-chip) is co-hybridized with the test-genome labeled in one color and a reference genome labeled in another color, and a hybridization-profile along the genome is read out as the normalized signal-ratio between the two probes.

35.2.1 FISH Analysis of Genomic DNA

Metaphase FISH Metaphase spreads are prepared from cells in culture, which become arrested in metaphase by treatment with colchicine, a drug that destabilizes microtubules (Figure 35.2). The culture in arrested state becomes swollen by hypotonic incubation and is dropped from some distance onto a glass slide, causing cells to burst while their chromosomes adsorb to the glass surface in groups representing the chromosomal composition of single cells. Hybridizing the probe of interest to this slide then allows us to spot at light microscopic resolution the respective genomic positions on the adsorbed chromosomes. As a matter of course, this approach is applicable only to dividing cells. Fixed biopsies or fresh tumor material that does not grow in culture cannot be investigated this way.

While diagnosing ploidies and polysomies is satisfactorily achieved by classical karyotyping, FISH adds the important option to record distinct genomic regions. Typical applications for metaphase-FISH are:

- physical mapping of genes and genomic marker-sequences: the resolution to separate two locations is limited to a few Mbp because of the dense packing of chromatin in metaphase; a more detailed view is possible, however, using protocols to spread the DNA (fiber-FISH);
- identification of site of ectopic DNA integration (e.g., upon stable transfection of cells);
- identification of genome homologies among species (zoo-FISH): hybridizing labeled whole DNA of one species to metaphase-spreads of another species readily reveals the extent of homology;
- detection of translocations via chromosome-*paints;* multicolor-FISH (see below) unravels unknown translocations;
- composition analysis of marker chromosomes via *painting* probes towards sections or whole chromosomes of interest.

Translocation: shift of a chromosomal region to a new position in the genome.

Marker chromosome: additional chromosome of unknown origin.

Multicolor FISH This takes advantage of the fact that genomic regions on metaphase chromosomes do not overlap. Thus, to separate signals from different probes, they may be

labeled by mixtures of fluorochromes (spectral signature). Three probes, for example, are readily distinguishable when labeled with colors A, B, and a combination of A and B, respectively. The ratio method extends this approach to the use of only two fluorochromes for labeling many probes, each in a well-controlled ratio of the two. Labeling all 22 autosomes plus X- and Y-chromosome of the human genome to allow their direct identification by color instead of spatial signature (i.e., size, length of p-arm, and banding pattern) is feasible by the combination of five fluorochromes only (multiplex (M-) FISH).

Interphase-FISH and Fiber-FISH As mentioned above, the preparation of metaphase spreads requires that cells do proliferate. However, valuable information is also obtained by FISH from interphase chromosomes of non-proliferating systems. Thus, numeric aberrations are directly accessible by the number of focal signals per cell nucleus. Ploidies, for example, are detected using repetitive probes towards centromeric regions. Though the chromosomal position of a signal is not directly visible, physical distances between targets are accessible, exploiting its linear relationship with the measurable mean squared distance. This approach yields even higher resolution (100 kb) than metaphase genome mapping since interphase chromatin is less condensed. In this respect even higher resolution down to 1 kb provides fiber-FISH, where chromatin becomes splayed by exhaustive spreading strategies (Halo-preparation). Fiber-FISH enables us to display on the gene level genomic aberrations like deletions, inversions, duplications, amplifications, or translocations.

35.2.2 Comparative Genomic Hybridization (CGH)

CGH (comparative genomic hybridization) was developed originally to analyze solid tumors for genomic imbalances. Its readout is the listing of genomic gains and losses in the genome of a particular cell population (e.g., tumor) relative to a reference cell population. There is, though, now less interest in this technology as modern high-performance sequencing has become quick and economic in answering the same question with higher accuracy.

CGH measures genomic imbalances while ignoring genomic aberrations like translocations and rearrangements, which do not affect the numerical balance (Figure 35.3). Furthermore, ploidies regarding the entire genome remain undiscovered, owing to the normalization of the data (see below).

The determination of genomic imbalances is of particular interest in cancer research to gain access to the causes of tumorigenic degeneration. Genomic loss refers to potential tumor suppressor genes, gain to proto-oncogenes. Comparing the gains and losses within huge collectives of patient-material allows us to narrow down minimal altered regions, so called *hot spots*, to be further analyzed by other methods, for example, to qualify candidate genes and to unravel functional pathways that are involved with tumor development.

CGH is a two-color ratio application. Test-DNA (e.g., DNA of a tumor biopsy) labeled in one color and reference DNA (from a healthy donor) labeled in another color are co-hybridized to a target (metaphase preparation: chromosomal CGH, immobilized array of characterized nucleotide sequences: microarray-CGH) and the relative intensities of the two signals is measured as a function of target localization. Sequences that are overrepresented in the test-DNA cause dominance of the test-DNA signal at the corresponding target site, losses are indicated by dominance of the reference-DNA signal. Since labeling efficiency between the two colors technically is not sufficiently reproducible, data-analysis requires normalization of the balanced state for each experiment.

Normalization Since genomic imbalances of an unknown tumor sample can be unexpectedly vast, it is difficult to identify a reference region within the measured dataset for normalization. In a simple approach one may accept that most of the genome is balanced and that normalizing the integral signal (test-signal/reference-signal set to 1) is sufficiently precise. Alternatively, there may be good reasons why a certain genomic region remained balanced.

Intermixture Imbalances measured by CGH appear weaker than their nominal value. This happens because of non-tumor cells, which are part of each biopsy and contribute a balanced fraction to the extracted test-DNA. Since dampening the ratios due to intermixture reduces the

Comparative Genomic Hybridization (CGH), Section 37.2.4

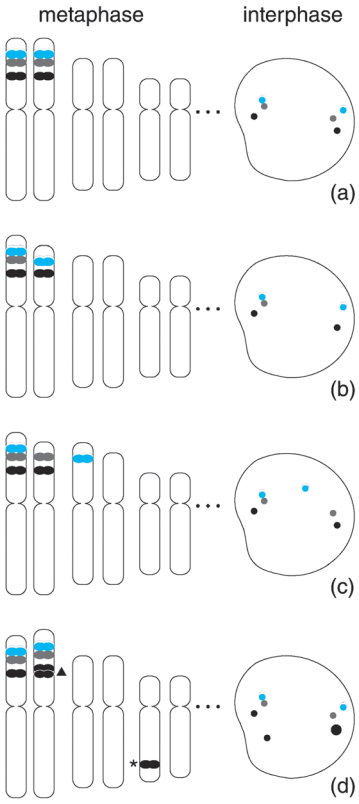


Figure 35.2 Scheme for common applications of metaphase and interphase FISH: Singular probes towards autosomal sequences regularly label two target loci, the maternal and paternal gene copy, respectively. In metaphase (left-hand side), signals appear paired since chromosomes consist of two sister chromatids, which is not the case in interphase. (a) Metaphase FISH towards three different target sequences directly demonstrates their sequential order, while interphase FISH only reveals neighborhood relationships upon statistical evaluation. (b) Both metaphase and interphase FISH allow direct demonstration of deletions. (c) To show translocations directly also requires metaphase spreads as a target. (d) Gain of genomic regions associated with translocation (asterisk) or polysomy become visible on metaphase as well as interphase preparations. Tandem repeats (triangle), however, are difficult to resolve spatially.

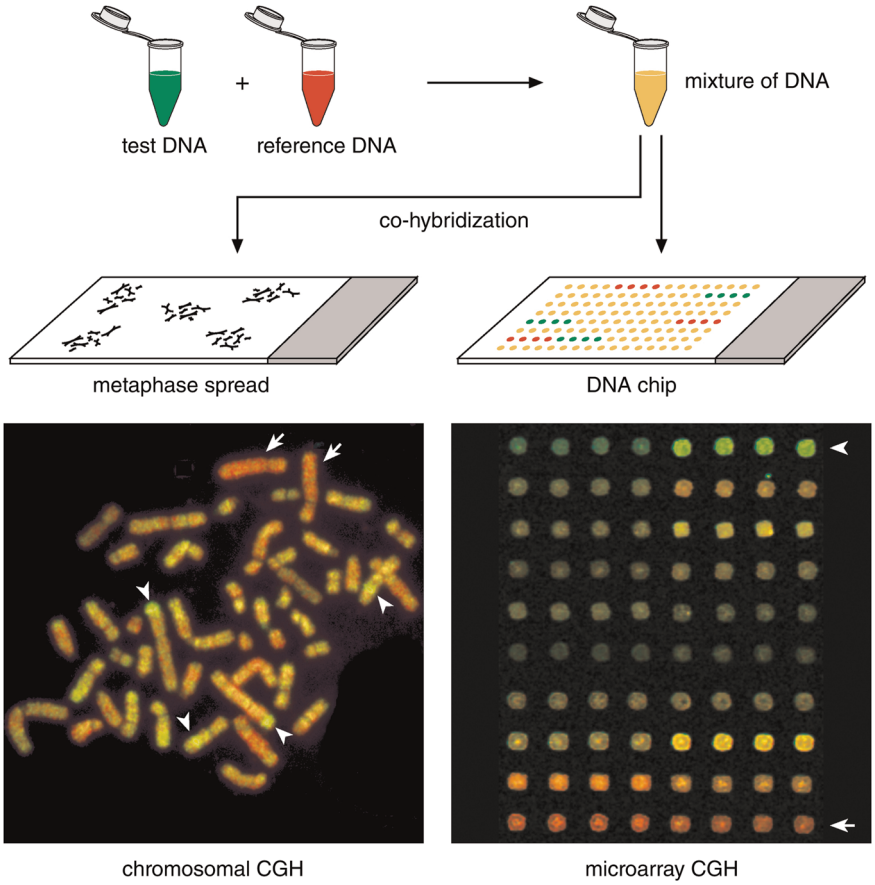


Figure 35.3 Chromosomal CGH and microarray CGH: CGH (comparative genomic hybridization) measures the ratio of fluorescence intensities of two probes, test and reference DNA, co-hybridized to the same target and distinguished by their reporter-fluorescence (here: green signal for tumor DNA; red signal for DNA of a healthy donor). Chromosomal CGH (left-hand side) uses metaphase spreads as targets, while the target in microarray CGH is an array of predetermined nucleotide sequences immobilized onto a glass slide (DNA chip). As a result (bottom), genomic gains within the test-genome are expressed by the dominance of green signals over red (e.g., arrowhead); the dominance of red signals over green indicates genomic losses accordingly (e.g., arrow).

sensitivity of the method, eventually only high-level amplifications are significantly detected. To improve the situation, cell sorting or micro-dissection techniques have been exploited to enrich fractions of tumorigenic cells before preparation of the test DNA.

Chromosomal CGH This is based on metaphase spreads as a target. Such spreads are typically obtained from cells of peripheral blood of healthy donors. Preparation methods are established and reliable. The target obviously represents the entire genome including all repetitive sequences, which need to be blocked before hybridization (CISS, see Section 35.1.4). After co-hybridization, samples become counterstained with DAPI, which allows the identification of chromosomes by a banding pattern reminiscent of classic Giemsa staining. Data acquisition and processing proceeds by the following steps: (i) Several metaphase-spreads are digitally imaged into three channels representing the test probe, reference probe, and chromosome structure. (ii) Chromosomes are identified by their structure (DAPI stain) and digitally segmented for each individual metaphase spread. (iii) Chromosomes are rectified by a type of computer-animation and superposed to obtain an averaged ideogram for the acquired dataset. (iv) Intensities for test and reference signals cumulated perpendicular to the chromosome axes are divided by each other, normalized, and read out as a profile for each chromosome ideogram (Figure 35.4). To show low-number gains they need to stretch more than 10 Mbp; significance for high-level amplifications is reached already for genomic regions shorter than this. Nevertheless, in searching for candidate genes the physical resolution of chromosomal CGH is rather poor.

DNA Microarray Technology, Chapter 37

Microarray CGH This technique (also known as matrix CGH or array CGH) uses DNA-chips as a target instead of metaphase preparations. These chips are glass slides with thousands of pre-chosen target sequences immobilized in an array of tiny spots. In contrast to metaphases, DNA chips do not necessarily cover a whole genome (tiling path resolution) and their composition of target sequences is a strategic decision. Different chips for different purposes are commercially available, for example, tailored for differential diagnosis of related disorders.

Microarray Comparative Genome Hybridization (CGH), Section 37.2.4

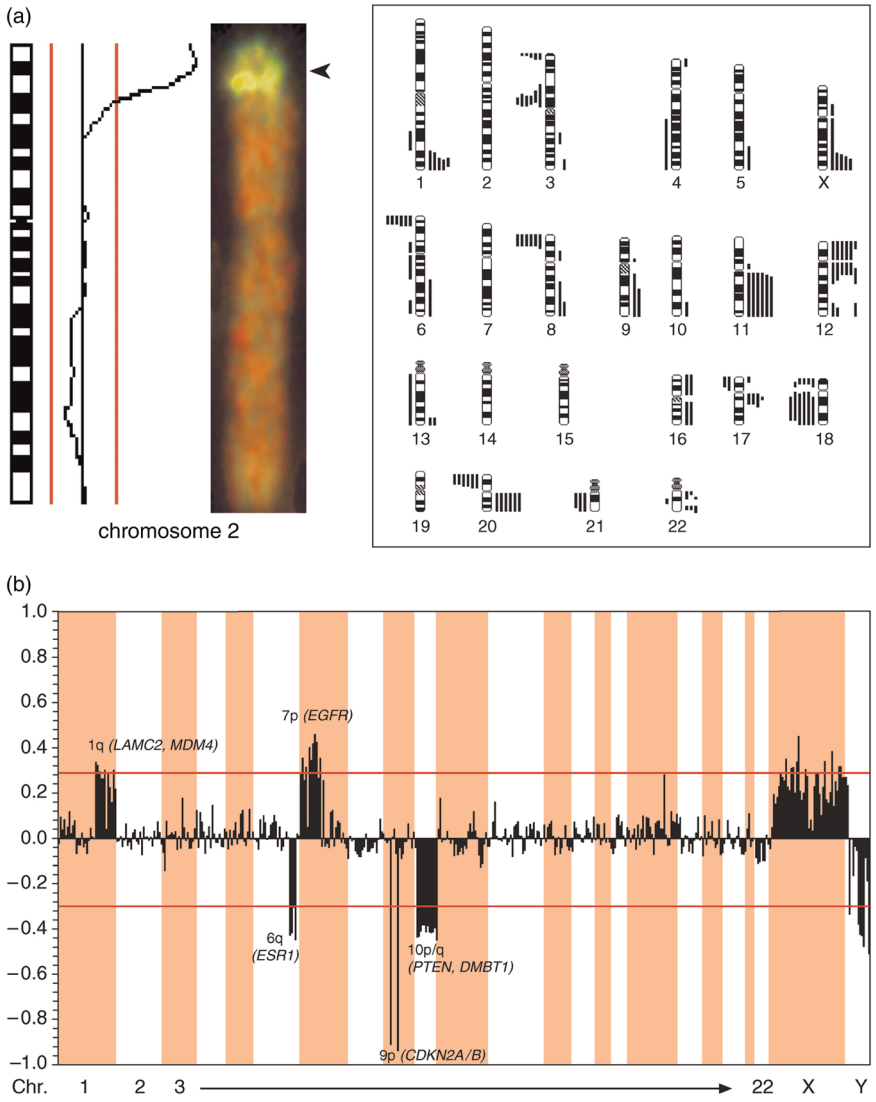


Figure 35.4 Exemplary demonstration of results from chromosomal CGH (a) and microarray-CGH (b). (a) On the left-hand side, gastric cancer with amplification of 2p23-p24 (arrowhead; see also Figure 35.3, which shows the entire metaphase spread of this experiment). The ratio profile plotted next to the chromosome 2 ideogram represents an average over more than ten single measurements. The nominally balanced state is marked with the black base-line; red lines to the left and to the right indicate the agreed confidence limit beyond which aberrations are scored. On the right-hand side in (a): To allow straightforward comparison among a whole patient collective, the extent of scored imbalances is visualized as bars allotted to corresponding positions in the ideogram (right, gains and left, losses). Recurrent aberrations within a collective of pancreatic cancer is readily appreciable in this example. (b) Bar diagram to visualize microarray CGH ratios measured for a case of glioblastoma (female patient, male reference-DNA): Each bar represents the CGH ratio of one target sequence plotted on a logarithmic scale in chromosomal order (1–22, X, and Y). Nominal value for the balanced state in logarithmic scale is null; ratios of balanced regions scatter around this baseline. Significance levels of aberrations can be calculated from the statistic mean variation of balanced regions for each experiment. Here, the confidence level of three times standard deviation is indicated by the red lines. Gain in chromosomal regions 1q23-q43 and 7p12, and losses in 6q24-q25, 9p21, and chromosome 10, are readily appreciable (in brackets: Candidate genes).

The chip-technology offers two important advantages relative to chromosomal CGH:

- The physical resolution is much better and limited only by the length of target sequence to obtain efficient hybridization, roughly 100 bp, which is short compared to the length of single genes.
- Analysis of microarrays is fully automatic and therefore capable of high-throughput approaches.

In the early days, DNA chips had been composed from DNA libraries. The management of such whole-genome libraries, including amplification, purification, characterization, and archival storage of the DNA, is tedious, while in practice the quality and assignment of ordered material was often insufficient. Nowadays, oligonucleotide-chips are used. The oligonucleotides become synthetized directly on-chip, or they are spotted from solution onto the chips by robotic loading. As they are artificially designed, they do not contain unwanted sequences, in particular no IRS. Hybridization signals are much higher compared to chips based on library DNA. Nevertheless, target-sequences are placed in replicates to increase the statistical significance of the hybridization experiment.

Hybridization and data acquisition are performed with special instrumentation to fit a workflow for high-throughput turnover, where false-proof transmission of sample data must be guaranteed. Data analysis involves huge program packages. Starting with two scanned

images of the chip representing the signal intensities for test and reference-DNA, respectively, the single spots are segmented by automatic image analysis and automatic filters applied to exclude spots of poor quality, for example, due to dirt, contact with their neighborhood, or weak intensity. After local background subtraction, the median of mean intensities from replicates is taken as the final measurement for each target sequence and exported in a table for normalization and segmentation of genomic imbalances with other software. In addition, further analysis relies on digital intelligence to be able to manage the enormous amount of data. Typically, one experiment delivers more than 100 genes as potentially disordered. Further criteria are required to filter for relevant clues in the data sets. Strategies span from elaborate cluster analysis to the screening of databases for pathway analysis.

Further Reading

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.

Gray, J.W., Pinkel, D., and Brown, J.M. (1994) Fluorescence *in-situ* hybridization in cancer and radiation biology. *Radiat. Res.*, **137**, 275–289.

Kallioniemi, A., Kallioniemi, O.-P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., and Pinkel D. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**, 818–821.

Lichter, P., Chang Tang, C.-J., Call, K., Hermanson, G., Evans, G.A., Housman, D., and Ward, D.C. (1990) High-resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. *Science*, **247**, 64–69.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, J., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, B.M., Gray, J.W., and Albertson, D.G. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Döhner, H., Cremer, T., and Lichter, P. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chrom. Cancer*, **20**, 399–407.

Trask, B.J. (1991) Fluorescence *in-situ* hybridization: applications in cytogenetics and gene mapping. *Trends Genet.*, **7**, 149–154.

Physical and Genetic Mapping of Genomes

36

Christian Maercker

Esslingen University, Kanalstraße 33, 73728 Esslingen, Germany

The creation of genetic and physical maps of genomes is an important area of research in the life sciences and medicine. Work in this field plays an essential role in understanding the genetics of organisms and enables prenatal diagnostics, diagnostics of complex diseases, and therapeutic approaches in personalized medicine. Scientists currently estimate over 10 000 diseases to be monogenic (WHO). In addition, many diseases have polygenic etiology, meaning that several gene defects contribute to the expression of the disease phenotype. One of the goals of gene mapping is to identify all genetic variants relevant for disease. A physical map contains the sequences of markers and shows the distances between markers within the genome (see cytogenetic methods described in the previous chapter). A genetic map is based on the analysis of the common inheritance of defined markers of known position coupled to a certain phenotype, in this case a specific disease.

Cytogenetic Methods, Section 36.1

36.1 Genetic Mapping: Localization of Genetic Markers within the Genome

Genetic information in eukaryotes is distributed on discrete chromosomes, which are visible during the metaphase stage of mitosis. An image of all the metaphase chromosomes within a single cell is described as its karyotype. The karyotype of a normal diploid human cell contains 46 chromosomes. Diploid organisms have two copies of each autosome and, in addition, two sex chromosomes. Each chromosome, in turn, consists of two complementary sister chromatids.

36.1.1 Recombination

Somatic cells proliferate by mitotic division. Before separation of the cells, cytokinesis, each chromosome is replicated and one copy of each chromosome is transferred into each daughter cell. The germ cells are generated by meiosis, meaning two successive cell divisions, resulting in a haploid chromosome set. During meiosis, genetic material is exchanged between non-sister chromatids in a process called crossing over, resulting in two recombinant chromatids. The probability of a crossover event between two loci on a chromosome is dependent on the distance between these loci. The recombination frequency is high for loci that are far apart, which results in them being inherited independently. Neighboring genes, on the other hand, are usually inherited together, exhibiting genetic linkage. Therefore, the degree of genetic linkage is an indirect measure of the physical distance between genes on a chromosome:

$$\text{Distance (cM)} = \left(\frac{\text{Number of recombinants}}{\text{Number of offspring}}\right) \times 100 \qquad (36.1)$$

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

One centimorgan (cM) is therefore defined as a recombination frequency of one percent. However, this type of genetic measurement is only possible over limited distances, since the probability of two crossovers increases with distance between two loci. In this event, two distant genes separated by two crossover events cannot be distinguished from ones where no crossover has taken place – thus the real distance between two loci on the genetic map might be underestimated. However, it is also possible to map longer distances by the use of intermediate steps, when possible. The distances between the respective single steps then are added to determine the genetic distance between the genes of interest. This method is also recommended to provide verification of distances measured directly. In the human genome, the genetic

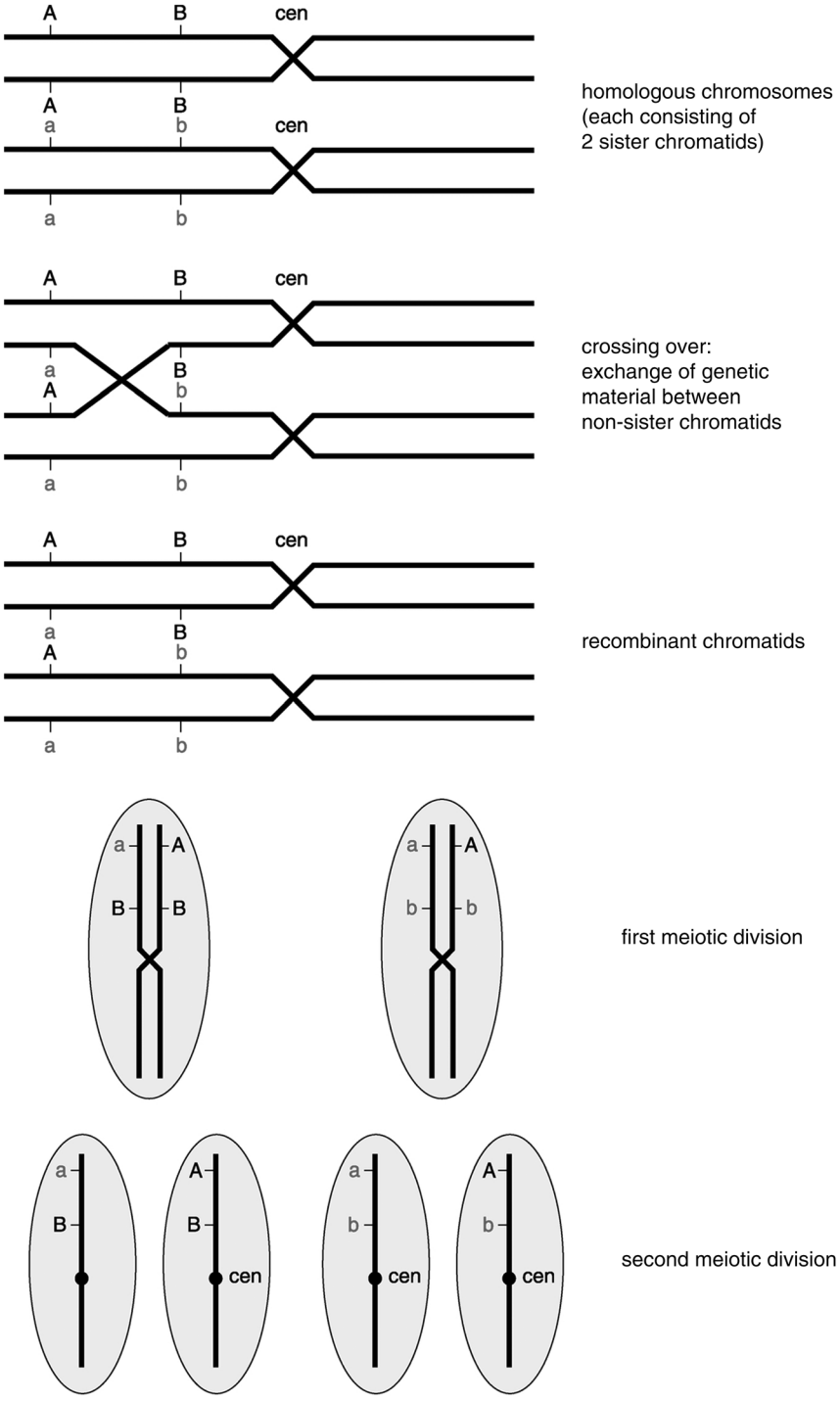


Figure 36.1 Recombination and crossing over. During meiosis, genetic material frequently is exchanged between non-sister chromatids. This results in recombinant chromatids containing maternal and paternal genetic material. Detecting recombination is only possible if maternal and paternal genetic material can be discriminated. This is possible via polymorphic markers that are differentially expressed. Here the marker at locus 1 appears as a or A, whereas locus 2 is detectable as b or B.

distance of 1 cM corresponds to about 1 Mb, when averaged over the entire genome. However, this correlation is of limited value, because the recombination frequency varies between different genomic regions and thus the perceived distance. Some regions are rarely affected by recombination events, whereas other regions show a very high recombination frequency and are thus referred to as recombination hot spots. Therefore, the determined genetic distance frequently differs from the calculated physical distance.

To be able to map the chromosomes quickly and completely, markers, fixed points on the genomic map, are determined. A genomic locus qualifies as a marker if it is present in different variants across the whole population. Only when parents carry different variants, alleles, of the markers, can a new combination in the F1 generation be observed. Such markers are called polymorphic markers (Figure 36.1).

The quality of markers is determined by their heterogeneity, which is described by the number of possible alleles and the relative frequency in the population. The more differing alleles a marker contains and the more evenly these markers are distributed among the population, the more helpful it is in analysis of the recombination events from one generation to the next. Markers that can be used for an individual analysis in this respect, that is, which appear in different alleles in both parents, are called informative.

Informative markers are defined points in the map of a genome in which different alleles are present in the parental generation, for example, microsatellite markers, a subset of the polymorphic sequence tagged sites (**STS**s).

Markers can be subdivided into two categories. For the generation of genetic maps, polymorphic markers are especially valuable, whereas for physical mapping a unique, well-defined position is most important – each marker must be attributable to a single copy at a single locus within the genome but they need not necessarily be informative. Genetic markers are restriction fragment length polymorphisms (RFLPs), microsatellites, and single nucleotide polymorphisms (SNPs)/single nucleotide variants (SNVs). Physical markers are genes, sequence-tagged sites (STSs), and chromosome breakage sites.

36.1.2 Genetic Markers

Restriction Fragment Length Polymorphisms Restriction fragment length polymorphisms (RFLPs) were the most important class of DNA polymorphisms for a long time. They are most often the result of single base changes, but can also result from insertions or deletions. If a change alters the recognition site of a restriction endonuclease, this leads to differences in the length of restriction fragments of the genome, which are easily observed. Although most sequence variations cannot be correlated to phenotypic changes, they nevertheless behave as Mendelian genes and can therefore be used as genetic markers (Figure 36.2).

Restriction Fragment Length Polymorphisms, Section 27.1.4

Microsatellites (Polymorphic STS) Microsatellites (polymorphic sequence tagged sites, STSs) are a special type of STS (Section 36.2.3). In the case of STSs, knowledge of the specific sequence of the locus is necessary. Conserved primer sites serve to amplify the polymorphic STS, while the sequences between the primer binding sites differ in length. Therefore, in contrast to non-polymorphic STS, allele-specific PCR products can be generated. These allelic differences are inherited according to Mendelian rules, which qualifies them as genetic markers.

In most cases, microsatellites consist of repetitive sequences with varying repetitions in the different alleles. These short repetitions appear in all eukaryotic genomes. Most repetitive sequences contain C/A units, which can be amplified by locus-specific flanking PCR primers. The number of C/A units is very often highly polymorphic, which is very convenient for genetic mapping. In addition, a high density of markers can be found, because the microsatellites can be as frequent as one repetitive sequence cluster per 50 kb of genomic sequence.

Single Nucleotide Polymorphisms (SNPs)/Single Nucleotide Variants (SNVs) SNPs/SNVs are variants of singe base pairs (bps) in a DNA strand. SNPs/SNVs represent about 90% of all genetic variants of the human genome and appear at different densities in certain genomic regions. In two-thirds of all SNPs/SNVs cytosines and thymines are exchanged, because in vertebrate genomes cytosine is very frequently methylated. By spontaneous deamination, 5-methylcytosine is converted into thymine. These exchanges are also generally called "successful" point mutations, that is, as genetic changes that have successfully spread to a certain degree within a gene pool of a population. In the International HapMap Project more than four million SNPs/SNVs were identified. Since SNPs/SNVs are very frequent and can be analyzed

Southern Blotting, Section 27.4.3

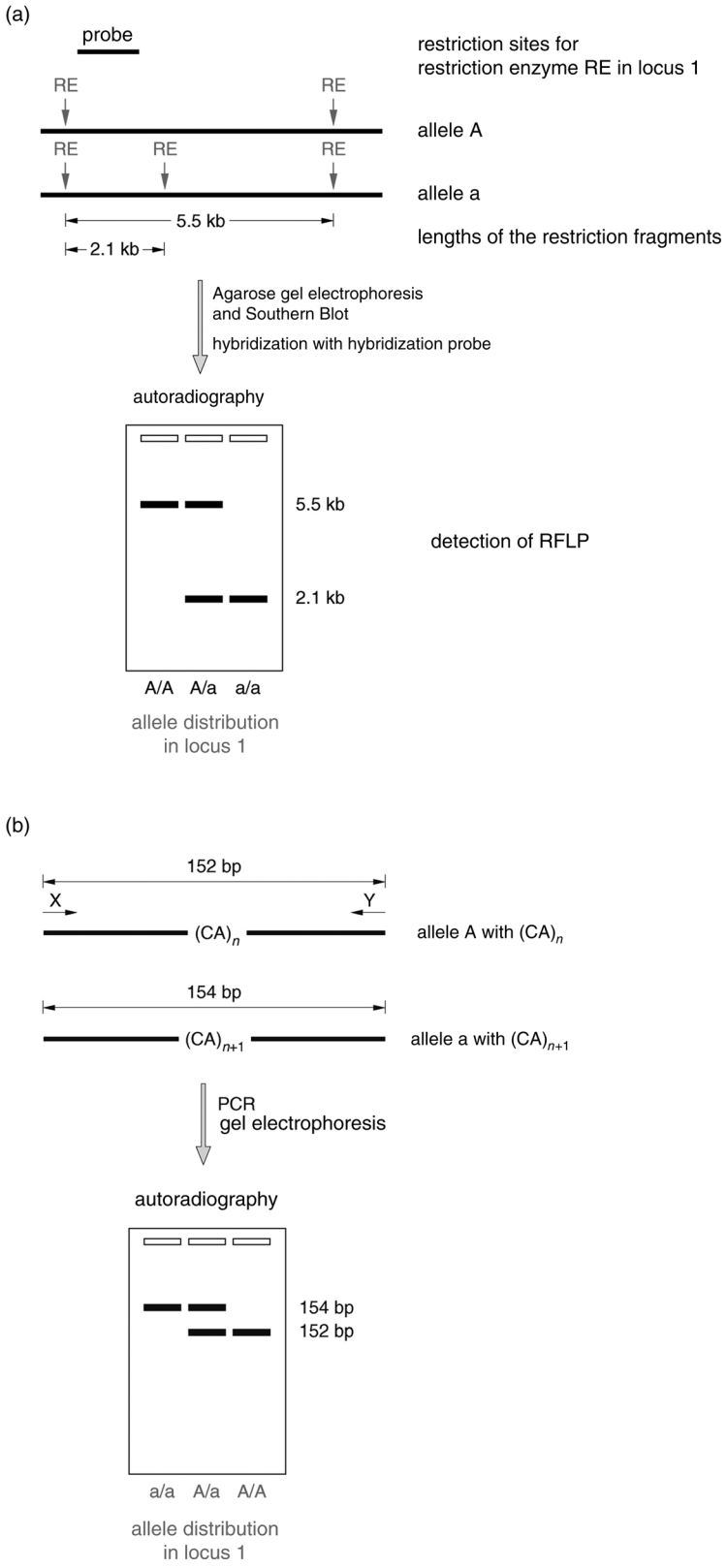


Figure 36.2 (a) Restriction fragment length polymorphisms (RFLPs). RFLPs can be recognized, when a restriction site (RE) is affected by a polymorphism. RFLPs can be detected with a hybridization probe that is located near the locus. The respective alleles can be detected by autoradiography. (b) Polymorphic STSs (microsatellites) are displayed by PCR, because the distance between the primer binding sites differs between different alleles. The variation between DNA fragments is often the result of a variable number of small repetitive elements, in most cases consisting of $(CA)_n$. For verification, genomic DNA can be separated by high-resolution gel electrophoresis (e.g., PAGE). This allows the detection of differences in length of even a single nucleotide.

using microarray analyses or high-throughput sequencing, they are now the most commonly used genetic markers.

DNA Micro Array Technology, Chapter 37

36.1.3 Linkage Analysis – the Generation of Genetic Maps

The generation of genetics maps is often called linkage analysis, because it investigates the common inheritance of two or more markers "linked" together. The neutralization of linkage happens by reciprocal exchange (double crossing over) of genetic material during meiosis (Section 36.1.1). Recombination between homologous chromosomes during meiosis is a common event. The resulting new combination of markers serves as the basis for linkage analysis. The goal is to find out if two loci are inherited more often than would be expected if they were located on two different linkage groups (chromosomes or distant segments of a single chromosome).

During meiosis, each pair of homologous chromosomes is independently subdivided into the daughter cells. One locus therefore co-segregates with a second locus on another chromosome with the probability of 50%. Loci on the same chromosome are expected to be separated by recombination with a probability of less than 50%, in relation to the distance between these loci. This percentage is called recombination fraction (cM) or θ, which is observed between two loci. The value of θ ranges from 0, for closely neighboring loci not separated by recombination, up to 0.5 for distant loci or loci located on different chromosomes. Thus, θ is a measure for the genetic distance between two loci.

Recombination fraction: proportion of recombination by crossing over between two markers during meiosis, which is in relation to the distance between the respective loci on the chromosome. The distance, given in centimorgans (cM), is determined by linkage analysis.

However, as previously mentioned, this principle is only valid for short distances. Since the probability of multiple recombination events increases with the distance between two loci, θ has to be converted into a genetic distance by a mapping function. Two loci are considered genetically coupled if θ is less than 0.5. The task of linkage analysis is to determine θ, and to calculate its statistical significance, when it is less than 0.5.

The χ2 Test The χ^2 test is a simple method for the detection of genetic linkage, which, however, is only valid for organisms with a large number of offspring. The statistical significance of a linkage can be estimated. As previously described, linkage analysis is based on the recombination frequency between two defined loci. This recombination frequency is reflected in the relative frequency of different classes of the products of meiosis.

Two heterozygous loci in a diploid organism (Aa Bb) can result in four different combinations in gametes: AB, ab, Ab, and aB. If both loci are not linked (not located on the same chromosome), all four possibilities are expected to be observed in the proportion of 1 : 1 : 1 : 1. If the markers are linked, the distribution is different and the marker combinations produced by recombination are under-represented. The first question to be answered is whether the proportion is 1 : 1 : 1 : 1. If this is the case, both loci are not linked. If the relation between the marker combinations is different, both loci are linked. If the deviation from equal distribution is obvious, mapping is simple. However, small deviations from the expected proportions have to be confirmed statistically in order to draw a clear conclusion.

An example: 500 products of meiosis are analyzed. The experimental results are as follows:

Class 1: AB 145
Class 2: ab 140
Class 3: Ab 105
Class 4: aB 110

The test of recombination frequency reveals $105 + 110 = 215$ recombinants (43%, $\theta = 0.43$). If inheritance of these markers is not linked the expected value is 50%, which differs from the observed value of 43%. Is the difference between the experimentally observed value and the expected value ($50\% - 43\% = 7\%$) significant or is it a random result, because only 500 meiosis products were analyzed? The χ^2 test is the method of choice to answer this question:

1. The null hypothesis posits no linkage.
2. The value of χ^2 is calculated. The number of test points is the critical variable for the evaluation of the significance of the results. For the calculation of χ^2, the number of

Table 36.1 Calculation of calculation of χ^2.

$$\chi^2 = \sum \frac{(N-E)^2}{\text{E of all classes}}$$

Class	N	E	$(N-E)^2$	$(N-E)^2/E$
AB	145	125	400	3.2
ab	140	125	225	1.8
Ab	105	125	400	3.2
aB	110	125	225	1.8
	500	500		$\chi^2 = 10.0$

experimental meiosis products (N) is compared to the number of meiosis products that would be expected according to the null hypothesis (E) (Table 36.1).

3. By means of the χ^2 value the plausibility, p, of the null hypothesis is calculated. The first step is to determine the degrees of freedom (df):

$$\text{df} = \text{Number of classes} - 1 \tag{36.2}$$

In our case: $\text{df} = 4 - 1 = 3$.

Table 36.2 shows the probability of the null hypothesis (grey-shaded values). With $\chi^2 = 10$ and $\text{df} = 3$ in our case, the probability of a non-linked inheritance for both analyzed loci is $p = 0.015$.

4. Acceptance or rejection of the null hypothesis. The value of $p = 0.05$ is usually chosen as the threshold valid for exclusion of the null hypothesis. Since the determined value of 0.015 is less than 0.05, the null hypothesis is rejected. The assumption is that a linkage exists in this case.

Tests for Plausibility For many pedigrees in higher mammals, including humans, it is often not possible to determine or measure all recombinants, as well as non-recombinants. Therefore, in these cases, instead of the χ^2-test, a method based on the calculation of the likelihood of recombination between two certain loci is applied. The calculation of the linkage probability

Table 36.2 Values of the χ^2 distribution. The values with a gray background apply to the example presented in the main text.

p	0.995	0.975	0.900	0.500	0.100	0.050	0.025	0.010	0.005
df 1	0.000	0.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319

between two markers is carried out with the aid of computerized programs. The resulting calculations of plausibility are placed in the formula:

$$Z(\theta) = \log_{10}\left(\frac{L(\theta)}{L(0.5)}\right) \tag{36.3}$$

Here L is the likelihood of a given θ. The L of $\theta < 0.5$ (linked inheritance) is compared to the L of $\theta = 0.5$ (non-linked inheritance). In linkage analysis, the result is often given as a logarithmic value of the quotient, also called LOD score ($Z(\theta)$; logarithm (base 10) of the odds).

Generally, we speak about loci being linked at a LOD score of 3 or more. At this LOD score, the likelihood that the loci are coupled is 20 : 1; analogously, a LOD score of 4 indicates a probability of linkage of 200 : 1. At a LOD score of 2 the statistical basis of the experiment should be extended by the investigation of further pedigrees in regard to the two loci investigated. Alternatively, neighboring loci can be included. If further individuals are investigated and the LOD score decreases, it can be presumed that the markers are not linked.

36.1.4 Genetic Map of the Human Genome

A genetic map consists of polymorphic markers – RFLPs, polymorphic STSs, and SNPs/SNVs – at a certain distance from each other. The primary benefit of a genetic map is that mutated loci responsible for certain diseases can be located in the genome by familial linkage analyses. This is usually possible without knowing the function of the gene.

Decades ago it was already possible to string the linkage of individual pairs of phenotypic (biochemical) markers, together with the determination of their sequential arrangement on the chromosomes, into a primitive genetic map. At present a "complete" map, covering the whole genome with a marker every 0.5–2.0 cM, has been completed (Figure 36.3, also see Section 36.2.3). Most markers are informative (Section 36.1.1) and clearly defined, such as polymorphic STSs (C/A repeat markers) and SNPs/SNVs. In the course of the human genome project, numerous C/A repeats within the entire genome were analyzed. Markers are important for the mapping of certain loci within the genome. SNPs/SNVs, originally identified within the

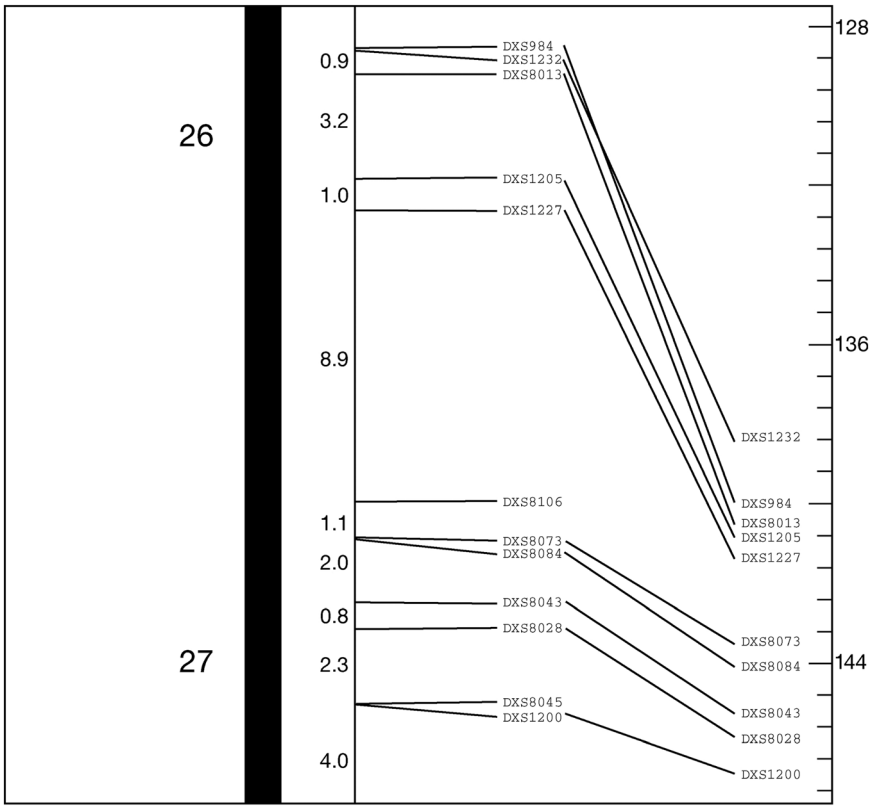


Figure 36.3 Physical and genetic map of the end of the long arm of the human X chromosome. On the left, the cytogenetic banding pattern is shown as an ideogram. The positions of the genetic markers are given in centimorgans. The physical distances between the same markers are given in megabases. The starting point for both measures is located at the telomere of the short arm. Source: After Nagaraja, R. *et al.* (1997) *Genome Res.*, **7**, 210–222. Copyright © Cold Spring Harbor Laboratory Press. CC-BY-4.0.

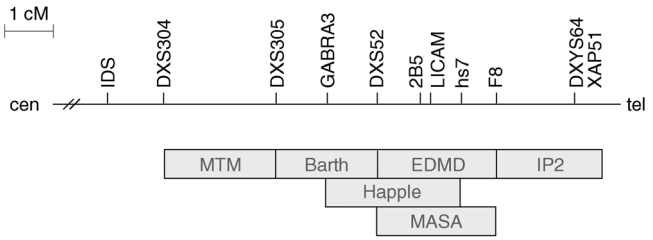


Figure 36.4 Candidate genes within the genetic map of Xq28. This simplified genetic map shows some genetic markers of the region that have been used to localize certain disease genes (codes above the line). The positions of genetic defects resulting in certain diseases are shown below. MTM: myotubular myopathy, Barth: Barth syndrome, EDMD: Emery Dreifuss muscular dystrophy, 1P2: *Incontinentia pigmenti* type 2, Happle: Happle syndrome, and MAS: MASA syndrome.

HapMap Project and later by high-throughput sequencing, are now the most important tool for the identification for genetic causes of disease.

36.1.5 Genetic Mapping of Disease Genes

A genetic map of high marker density is of great value, since it allows the mapping of genetic diseases within short time periods, if a pedigree of a sufficient size is available. Classical genetic mapping often took years to map a single gene. The mapping data, together with the DNA sequence information (physical map), are an important prerequisite for positional cloning (Section 36.2.4) and isolation of the genes involved in disease (Figure 36.4).

Many very common and dangerous diseases have polygenic causes, such as cardiovascular diseases, cancer, and schizophrenia. They are referred to as multifactorial diseases, because they are caused by the interplay between genetic and epigenetic aberrations, as well as environmental influences. Genetic predisposition influences the sensitivity to extra-genetic risk factors. The investigation of individuals with an identical genetic predisposition enables the characterization of the susceptibility of an individual to extra-genetic risk factors and offers a promising approach towards preventive personalized medicine. Personalized medicine involves adjusting medications according to the genetic background of the patient.

Genetic predisposition: mutations in certain genes, which determine sensitivity of an individual to extra-genetic risk factors.

The quality of mapping or LOD scores, respectively, of multifactorial diseases is limited by the size of the pedigree and marker density. The analysis of a possible genetic disease includes as many generations and affected relatives as possible. The DNA of all individuals is typed and linked inheritance is determined with the help of genetic markers. The linkage is analyzed via probability tests, as used for the generation of a genetic map (Section 36.1.3). Again, the quality of localization of the disease is dependent on the size of the pedigree and how polymorphic the markers are. The locus is usually not located precisely but instead is described as lying between two known genetic markers (Figure 36.4).

Genetic mapping is often just as important to exclude genes as candidates for involvement in certain diseases as it is to include them as possible candidates. The phenotype of an affected individual is very often correlated with certain genes, such as, for example, for certain oncogenes frequently being mutated in cancer. If an oncogene is causal for a disease, no recombination is detected by linkage analysis between a marker for the disease phenotype and the candidate gene locus. This, however, does not mean that the same oncogene is relevant in all patients suffering from the same type of cancer, because different mutations, even in differing genes, may lead to the same phenotype.

36.2 Physical Mapping

Physical mapping includes all methods that describe the distance in nucleotides between two individual markers within the genome, as opposed to genetic maps, which are based on recombination frequencies between markers.

36.2.1 Restriction Mapping of Whole Genomes

One possible way to map chromosomes or a complete genome is the description of the order of the restriction fragments (top-down procedure, Figure 36.5a). After restriction mapping, certain regions can be cloned. This method was usually applied to small genomes such as *Escherichia*

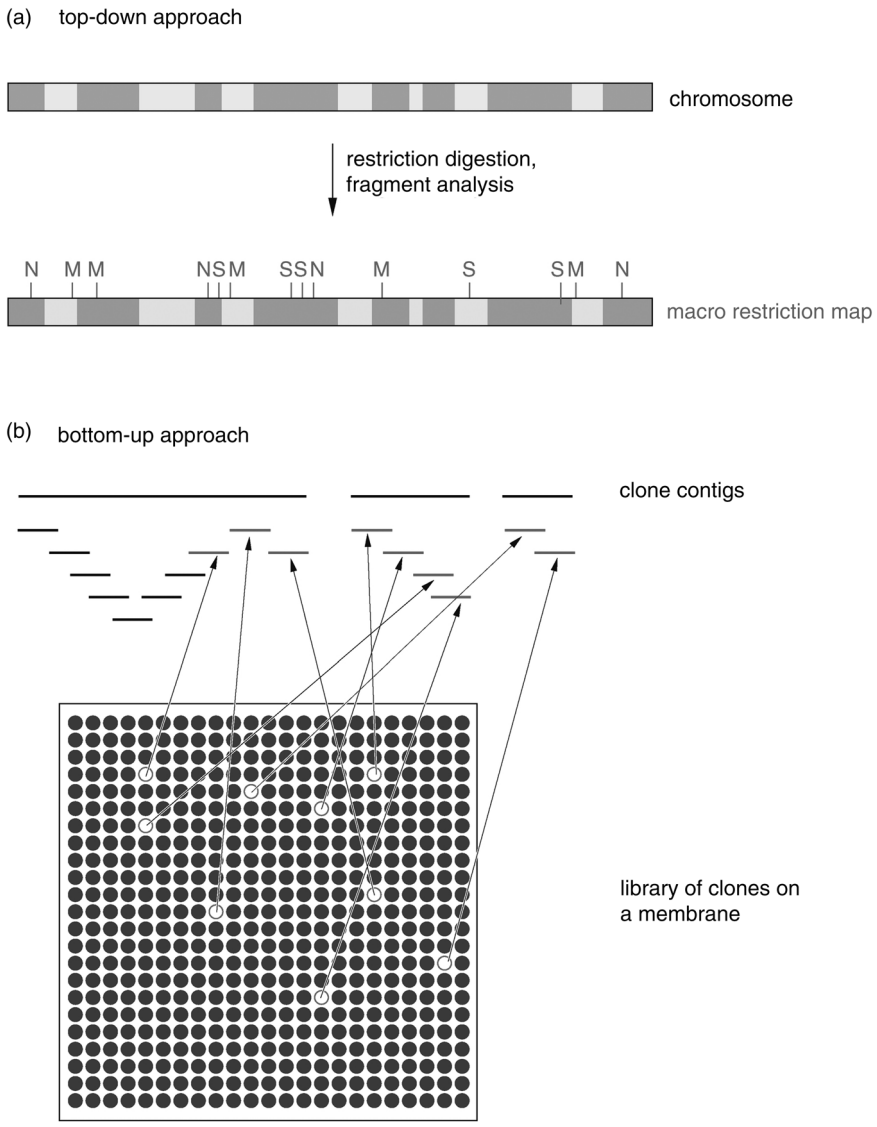


Figure 36.5 Two techniques of physical mapping. (a) Top-down approaches start from a cytogenetic map or a restriction map of genomic DNA or a single chromosome. Here, the restriction sites of rare cutting restriction enzymes, which generate large DNA fragments, are shown (N, NotI; S, Spl1; M, MluI). Since the genomic DNA was cloned prior to restriction digestion, "complete" maps arise from this procedure. These maps have a low resolution (>100 kb). Therefore, it is not possible to directly isolate clones or genes, respectively. (b) Bottom-up methods start with the generation of a library of clones, from which clones are isolated. Overlapping clones are identified by chromosome walking and combined to contigs (continuous sets of clones). These contigs lead to a high-resolution map, which, however, in most cases does not cover a whole chromosome (partial maps). Therefore, both methods are often used in combination.

coli or viruses. Type II restriction endonucleases are used for restriction mapping. These enzymes recognize specific DNA sequences, often, but not always, consisting of 6 bps, and cut the DNA within these sequences. Enzymes that cleave rarely within the investigated genome are favored because they produce long DNA fragments after restriction digest. The 4.5 Mb genome of *Escherichia coli* was separated into 21 DNA fragments with *Not*I, a restriction enzyme recognizing an 8 bp sequence. This allowed the sequential position of the DNA fragments within the circular genome to be determined.

Known genes or markers, respectively, can be assigned to certain restriction fragments by hybridization. If two markers are located on one restriction fragment, the physical distance between these two markers cannot be greater than the length of the analyzed fragment. For this type of physical analysis, the genomic DNA is cleaved with a restriction enzyme, separated by pulsed-field gel electrophoresis (Section 27.2.3), blotted onto a membrane (Section 27.4.3), and hybridized to a cloned genomic DNA fragment. This type of restriction mapping gets more and more difficult with increasing complexity of the genome, because even infrequently cutting enzymes like *NotI* generate many restriction fragments of similar length that cannot be effectively separated and individually analyzed.

In higher eukaryotes, cell hybrids allow generation of limited restriction maps with lengths of up to several megabase (Mb). Several of these partial maps can then be integrated to complete maps. For radiation hybrid mapping, human-mouse or human-rat cell hybrids are used, each

Restriction Analysis to Certain Restriction Fragments, Section 27.1

Pulsed-Field Gel Electrophoresis, Section 27.2.3

Southern Blotting Blotted onto a Membrane, Section 27.4.3

containing a limited amount of the human genome. A portfolio of about 100 different cell hybrids, each containing a different – possibly overlapping – set of human chromosomes, allows a mapping accuracy of 1–5 Mb. Since the resolution of this mapping method is limited, pocket maps (bins) are generated instead of conventional linear maps. More detailed information is possible by detailed physical mapping (see below). Since whole genomes can now be decoded by high-throughput sequencing within a single day (Section 36.2.3), restriction maps are no longer of much importance.

36.2.2 Mapping of Recombinant Clones

For restriction mapping of whole genomes, as described in the previous chapter, native genomic DNA is used. Multiple copies of the entire genome are used in the analysis, such that all loci are represented at an identical copy number. Cloning the genomic DNA is required for further analysis and manipulation of individual genes or gene segments. Not all genomic regions are equally accessible to cloning, which makes it a challenge to achieve complete coverage of the genome. For example, to clone 1 Mb genomic DNA into a DNA vector cloning system that can hold up 20 kb per clone, 500 clones are needed for a tenfold coverage ($500 \times 20\,\text{kb} = 10\,000\,\text{kb}$). The number of clones (N) that are necessary to clone a certain locus with a defined probability can be calculated:

$$N = \frac{\ln(1-p)}{\ln\left(1-\frac{1}{n}\right)} \tag{36.4}$$

Here p is the probability that a certain locus will be contained in a genomic library. The variable n describes the relation between genome size (here: 1 Mb) and the average insert size of the cloning system (here: 20 kb). If the goal is to clone each part of the genome with a probability of at least 99%, it results in the following equation:

$$N = \frac{\ln(1-0.99)}{\ln\left[1-\frac{1}{\left(\frac{1000}{20}\right)}\right]} = 228 \tag{36.5}$$

Therefore, 228 clones are necessary to clone each locus of the genome with a probability of 99%. The equation also shows that the number of clones to be analyzed is dependent on three parameters: target probability, genome size, and average insert size of the used cloning system (DNA vector used for cloning). However, the cloning system is the only variable, because the genome size of the organism is constant, and the quality of the library correlates with the probability that one clone can be isolated from this library. Moreover, the calculations described above are based on the estimation that all regions can be cloned with the same efficiency, which is not what has been observed for any organism analyzed so far. Therefore, in practice, it is necessary to clone much higher numbers than calculated.

Cloning Systems for Genomic Libraries For the generation of genomic libraries, different cloning systems are in use. The cloning vectors differ in capacity (maximum insert size that can be cloned), reproduction of cloned DNA (copy number per cell), stability of insert DNA (recombination rate), and manageability (DNA isolation, accessibility for DNA sequencing). A summary of the characteristics of cloning systems in use is shown in Table 36.3.

Aside from yeast artificial chromosomes (YACs), all DNA vectors use prokaryotic hosts (usually *Escherichia coli*). Decisive for the size (clone number) of the library is the cloning capacity of the DNA vector, since the number of clones representing the whole genome decreases with increasing insert size. This may ease genome analysis; however, the quality of clones is also critical. In eukaryotic YAC vectors, a large number of clones undergo rearrangement by recombination. Depending on the library, 25–70% of the clones are affected, which is a big disadvantage for analysis. Moreover, YAC DNA has to be purified by pulsed-field gel electrophoresis (PFGE), which is very labor-intensive. In contrast, isolation of P1 DNA (artificial phage P1 chromosome) and bacterial artificial chromosomes (BACs) in high quality and quantity is much easier. The inserts in these cloning systems are very stable, because they are propagated in modified *E. coli* hosts with a partially inactivated recombination system.

Table 36.3 Common cloning systems.

	Host	Insert size (kb)	Copies per cell	DNA-isolation[a)]	Direct sequencing	Rearrangement of the inserts[b)]
Lambda	*Escherichia coli*	5–25	>250	Good	Good	Very rare
Cosmid	*Escherichia coli*	35–45	3–50	Excellent	Very Good	Possible
P1	*Escherichia coli*	70–100	1–2	Very Good	Good	Rare
PAC[c)]	*Escherichia coli*	70–300	1–2	Very Good	Good	Rare
BAC[d)]	*Escherichia coli*	50–300	1	Good	Good	Very rare
YAC[e)]	*Saccharomyces cerevisiae*	50–2000	1	Difficult	Difficult	Frequent

a) Simplicity of DNA isolation and resulting purity.
b) Proportion of chimeric clones; frequency of deletions of the inserts.
c) P1 artificial chromosome.
d) Bacterial artificial chromosome.
e) Yeast artificial chromosome.

The bottom-up procedure for physical mapping starts with the generation of a genomic library. To generate a contiguous set of clones, referred to as a contig, the cloned fragments are sorted according to overlapping regions by hybridization (Figure 36.5b). Sequencing and computer analysis allow construction of a continuous DNA sequence of the genome, provided it is represented by the genomic clones under analysis.

The shotgun strategy circumvents the sorting of clones of a genomic library. Instead, genomic DNA is fragmented by restriction enzymes or other means, cloned into plasmid vectors with small insert sizes (about 2 kb) and randomly sequenced. Computer analysis then allows the identification of overlapping clones to produce a continuous genomic sequence. Newer methods do not even require cloning, because the genomic DNA fragments are directly amplified and sequenced by high-throughput methods (Section 36.2.3 and Chapter 30).

36.2.3 Generation of a Physical Map

Using the clone-based strategy, the clones containing genomic DNA are ordered along the chromosomal DNA in sequence (Figure 36.5). The clones are sorted similarly to the pieces of a puzzle, only in this case the pieces (genomic DNA inserts of the clones) overlap. Since it is generally necessary to work with a huge number of clones, a quick and unambiguous method is necessary for mapping (see below). However, since many genomic sequences are accessible in public databases, physical maps of many genomes are readily available.

STS Mapping In the 1980s, the concept of sequence tagged sites (STSs) was developed. These are 100–300 bp DNA sequences that are unique within the genome. The DNA pieces are amplified by PCR and used as markers for physical mapping. The PCR product describes a unique locus within the genome. In principle, each genomic locus is qualified for a STS sequence. First, the sequence is aligned with all sequences in databases to make sure that the primer sequences are not located within repetitive sequences. It does not matter if the region between the pair of primers is repetitive (Figure 36.2b).

For physical mapping it is not necessary that the employed STSs are polymorphic, only the amplification product has to be unambiguously detectable by PCR. Possible STS markers are DNA segments of unknown function, polymorphic DNA markers, and genes (Section 36.1.2).

Yeast artificial chromosome (YAC) libraries can be used for chromosome mapping, distributed as one clone per well in microtiter plates. PCR is used to determine which YAC clones can be amplified with a STS primer pair. Since YAC libraries consist of several thousand clones and it therefore would be very labor-intensive to individually test each YAC for each STS, special methods have been developed to allow efficient screening. Pools containing DNA from several clones can be pooled into superpools of up to seven levels of smaller pools. By intelligent combination of YACs it is therefore possible to find the possible YAC clone with only a few PCR reactions (Figure 36.6).

Techniques for the Hybridization and Detection of Nucleic Acids, Chapter 28

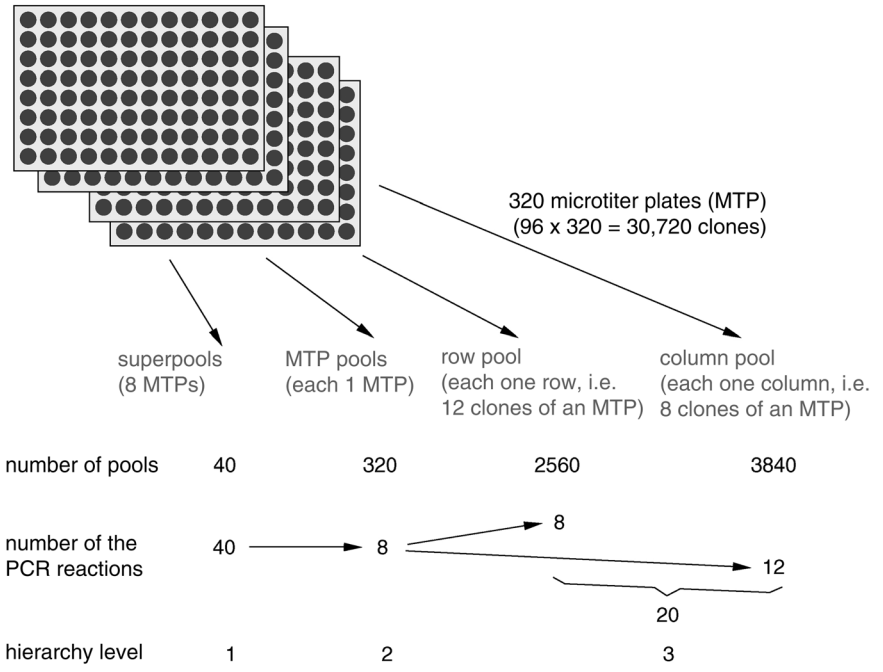


Figure 36.6 STS screening. Here, a YAC library of the human genome is shown, consisting of 30 720 clones, stored in 320 microtiter plates (96 clones per plate). Pool creation: In the three-step pool system shown in this figure, DNA samples from all clones are isolated in a central laboratory. The DNA of each clone is mixed with DNA from other clones of a library in different combinations (pools). The superpools contain DNA from eight microtiter plates (8 × 96 clones). In our example, this results in 40 superpools. Microtiter-plate pools contain DNA of all clones from a single plate (here: 320 microtiter-plate pools). The pools in rows and columns contain DNA of the clones of one row (12 clones) or one column (8 clones), respectively. Hierarchical screening: In the first step, the 40 superpools serve as samples for the STS-PCR. By PCR and gel electrophoresis, the positive superpools are identified, followed by STS-PCR of the microtiter-plate pools, which make up the positive superpool (8), in order to identify the microtiter-plate of the positive clone. In the third round of screening, the twelve row pools and eight column pools of the positive microtiter-plate are used for the amplification to find the precise clone responsible for the positive response.

Mapping by Hybridization to Cloned Libraries Physical mapping by hybridization of cloned genomic DNA is, in most cases, based on reference libraries. The principle is shown in Figure 36.7. Reference libraries are genomic libraries whose clones are sorted into microtiter plates. Each clone has an individual physical address and therefore is accessible individually. The clones are densely spotted on nylon membranes in an ordered and reproducible pattern. The "master filters," produced by only one or a few reference laboratories, can then be provided to individual working groups worldwide. This allows mapping experiments to be carried out in different laboratories with the same primary material, which makes it easier to compare and integrate the data. Moreover, information about genome libraries has been collected into databases. The starting point of such an experiment is the hybridization of a gene probe to nylon membranes containing genomic clones. With this screening approach it is possible to identify a set of clones representing a whole genome. This approach also is applicable for cross-hybridizations between evolutionary conserved genomes (e.g., human and mouse).

Oligonucleotide fingerprinting uses a series of short, 8–12 bp long, oligonucleotides that are hybridized to spotted membranes in numerous cycles. With this approach it is possible to assign a clone to a specific group of clones and eventually to a single clone, which is then added to a library.

Hybridization of libraries and mapping with the help of STSs are two techniques based on complementary principles to physically map a genome with recombinant clones. The first approach uses the clones themselves as references, whereas the second approach is based on the known information about STSs. In many cases, both methods have been used in combination, that is, reference clones were partially sequenced to generate STS markers, and the resulting STS amplification products were employed as probes for the next round of hybridization with reference libraries.

The clones of many BAC and PAC libraries have been sequenced and are publicly available. Since the sequences have been localized to the chromosomes by the minimal tiling path approach, the specific application of genomic clones (e.g., FISH (fluorescent *in situ* hybridization), BAC arrays; see below) is possible without the need for an elaborate screening to isolate the clones.

Mapping by High-Throughput Sequencing Clone-based mapping procedures are being replaced by direct sequencing without the need to clone the genomic DNA. This is made possible by next-generation sequencing. Most of these high-throughput methods start with a library

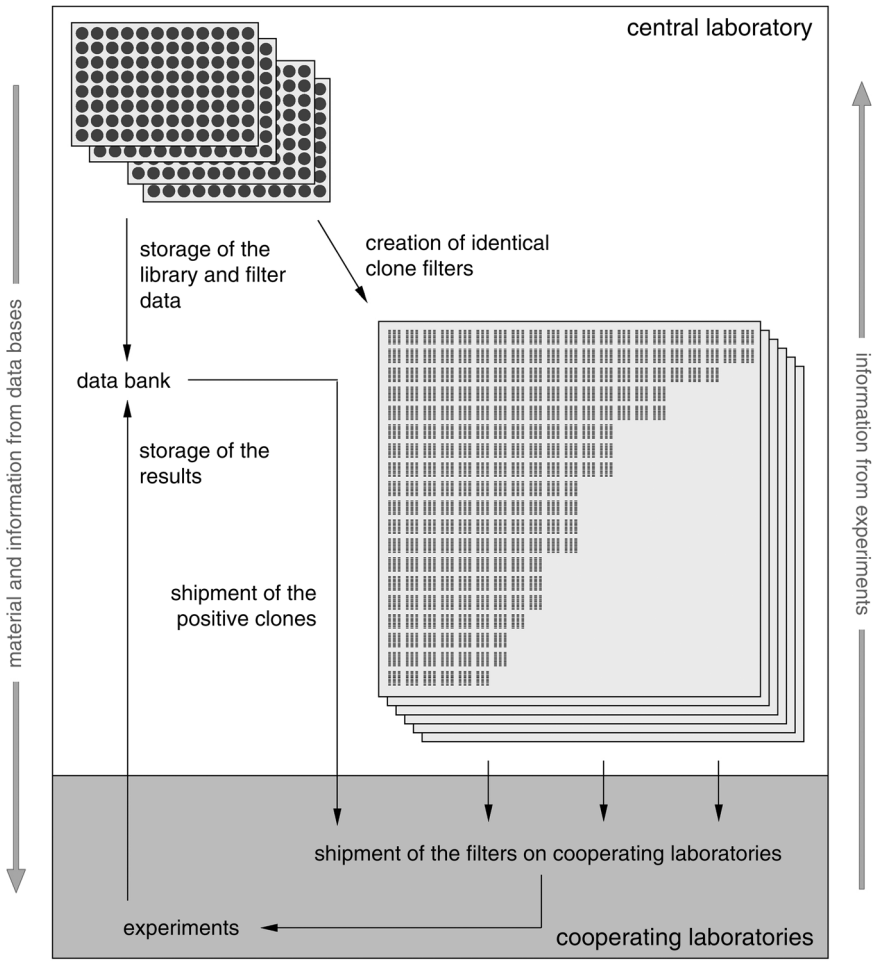


Figure 36.7 Reference system. The libraries used for mapping and gene isolation are provided by a central laboratory. The clones of the library are stored in the wells of microtiter plates. Clones can be transferred reproducibly and at high density onto nylon membranes using robots. These spotted membranes can be used by cooperating laboratories for hybridization. The coordinates of the hybridization signals can then be sent back to the central laboratory, which then sends the positive clones to the cooperating laboratory. Since the experimental results can be stored in public databases, laboratories all over the world have access, allowing for less redundancy of experiments. This practice has not only been successful for human material, but also for model organisms like mouse, rat, *Drosophila*, and yeast.

generation step, in which genomic DNA is fragmented by ultrasound or nebulization, ligated to standard oligonucleotide primers, and amplified by highly parallelized PCR in oil–water emulsion, or starts from immobilized DNA templates in array format. Each amplification product can then be individually sequenced in parallel. The sequences, with lengths of up to several hundred bases, are aligned by computerized algorithms to generate a complete genomic sequence. The alignment takes advantage of many sequences of different organisms already available in databases. Alternative approaches are sequencing by ligation or single-molecule sequencing.

36.2.4 Identification and Isolation of Genes

Since the invention of recombinant DNA technology, several thousand genes involved in disease have been isolated and analyzed. Typically, genes have been identified that lead to a certain disease due to a specific biochemical defect. At first, known protein sequences or protein sequences derived from nucleic acid sequences contributed to the identification of the genes. In some cases, antibodies against particular proteins allowed the isolation of a gene from a protein expression library. With some transformed oncogenes, the function could be used directly to identify the gene. All these techniques are summarized as functional cloning, since the gene product or the gene function serves as basis for the isolation of the gene.

The phenotype of a disease can, however, only very rarely be connected to a single protein or protein function. Hence, other strategies are necessary to isolate genomic regions responsible for a certain phenotype. Technologies have been developed that use the position of the gene as the basis for gene isolation. The specified gene locus is the key information and the isolation of

Positional cloning (reverse genetics) The cloning of a gene transcript of known position within the genome for use in functional studies.

the transcript is based on exact mapping within the genome. Originally, this approach was called reverse genetics; today the name positional cloning is more common. The latter more precisely describes the method, which starts with the position of the gene within the genome and leads to the description of the protein via the transcript. Other methods used to determine the possible functions of genes are described below.

Prediction of Gene Characteristics – Candidate Genes By exact investigation and monitoring of patients, in some cases it has been possible to describe gene defects responsible for a certain disease. These predictions concern the potential functions of the affected gene, which might be directly linked to the disease. Tissues or organs that are affected by a disease might allow conclusions about cell-type specific gene expression. Diseases linked to developmental disorders are most likely caused by genes expressed during certain developmental stages. A hereditary disease with increasing prevalence, or anticipation, from one generation to the next makes conclusions about the mechanism of mutation possible, which allows predictions concerning single parts of the gene sequence. If all these criteria are considered, the number of candidate genes can be narrowed, sometimes to the point where only one or a few remain as candidates. In the past, this method allowed the identification of genes associated with disease without any mapping information.

Gene Modeling Based on the genomic sequences, it is possible to predict the functions of genes. In lower eukaryotes such as, for example, yeast (*Saccharomyces cerevisiae*) the genomic sequence is mostly collinear with the translated, expressed sequence, that is, the genes only rarely are interrupted by introns. Therefore, after decoding of the complete genome, it was possible to determine open reading frames (ORFs), which are potentially protein encoding regions, within the yeast genome. Afterwards, the prognosticated gene activities were shown experimentally. By this approach, more than 5000 yeast ORFs could be verified. The smallest prokaryotic genome known so far is the genome of *Mycoplasma genitalium*, which contains a mere 470 genes.

The prediction of gene functions is much more difficult in higher eukaryotes, because their primary transcripts consist of exons and non-coding introns. The identification of exons in genomic sequences is often difficult, because the average length of one exon only is about 150 bp. Some exons are as short as 15 bp.

Single exons can be predicted with the help of computer programs that are based on neural networks. The reliability of these programs increases with the number of known genes that can be used as training set. Therefore, the predictions based on genomic sequences increases with new genomic sequencing data. However, it is still difficult to recognize complete coding gene sequences based on the genomic sequence, since computer analysis often misses exons or predicts them incorrectly. Moreover, terminal non-coding regions might not be recognized by the computer programs.

Therefore, it also is important to identify exons experimentally, using, for example, exon trapping. Here, certain chromosomal regions are cloned into a DNA vector that contains a "minigene" consisting of two exons and the associated splice acceptor (SA) and splice donor (SD) sites. If a chromosomal region containing complete exons and introns together with SA and SD sites is integrated into a DNA vector, the splice pattern of the transfected cells changes, which can be proved by PCR.

Positional Candidate Genes This approach to identifying disease genes combines the strengths of the determination of gene characteristics with the exponentially increasing sequencing data of regions with no known function. In the positional candidate approach, all available data about a genetic disease is used to learn as much as possible about the candidate gene. These gene predictions are compared with the genetic mapping of the disease gene. This means that all genes that are located in the genetic locus responsible for the disease are investigated for suitable characteristics. The idea is that characteristics of a disease that was mapped to a certain part of the genome are connected to genes located in this region (Figure 36.8). Candidate genes can then be investigated in patients for the presence of corresponding mutations. The success rate of this approach is proportional to the amount of sequence information available, which is steadily increasing and, thus, so are the prospects for its success.

It is not enough to detect the expression of ORFs in order to recognize defects in gene expression. The sequences responsible for gene regulation (e.g., promoters, enhancers,

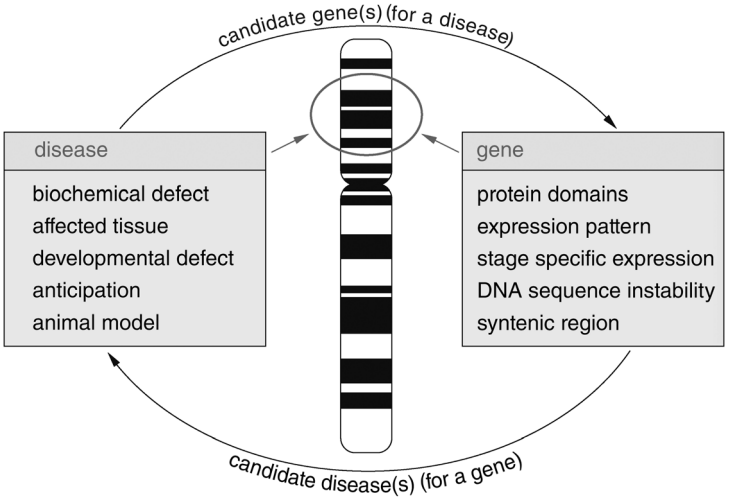


Figure 36.8 Positional candidate gene approach. A disease that has been mapped to a specific region of a chromosome can be assigned to a list of genes, which have been physically localized to this chromosomal domain. The characteristics of these candidate genes can be compared to the known properties of the disease. A direct correlation of disease and gene is sometimes possible. Potential biochemical defects of a disease can be compared to protein domains encoded by the gene (Kallmann syndrome). The affected tissues or developmental stages can be correlated with the expression pattern of certain genes (X-chromosome agammaglobulinemia). If symptoms are increased from one generation to the next (anticipation), they are often correlated with instable DNA sequences (myotonic dystrophy). If nothing else, knowledge about an analogous disease in model animals can serve as basis for a search in the human genome. Regions in which genes are represented in the same order in different organisms are called syntenic (homologous) regions. Source: figure modified after Bick, D.P. and Ballabio, A. (1993) *Am. J. Neuroradiol.*, **14**, 852–854.

silencers) also have to be identified. The sequences themselves, as well as their localization, are very heterogeneous. Therefore, their identification based on the genomic sequence is difficult. One promising experimental approach is ChIP chip (chromatin immunoprecipitation and chip analysis). Chromatin fragments, consisting of genomic DNA as well as proteins and RNAs bound to this DNA, are immunoprecipitated with antibodies against, for example, transcription factors. Afterwards, the DNA part of the precipitated complex is isolated and hybridized to a DNA microarray covering the whole genome (tiling array). The hybridization signal then corresponds to the DNA sequences bound to the respective transcription factor. Alternatively, the DNA sequences can be identified by high-throughput sequencing (ChIP-seq).

36.2.5 Transcription Maps of the Human Genome

Only about 1% of the human genome is transcribed into mRNA. Starting from the 1990s, the number of transcription maps showing expressed genes that have been localized to the genome has been steadily increasing. Physical mapping of expressed sequence tags (ESTs) has been important in this respect. Originally, mainly specific transcription maps covering only a few hundred kilobase up to several megabase were created. Now transcription mapping is performed genome-wide.

Sequence comparison is used to map cDNA sequences to the genomic sequence. However, isolation of cDNAs was not always successful for many ESTs, which are only expressed under certain physiological conditions, at or in low copy numbers, or with a short half-life. Since the advent of high-throughput methods, transcripts are normally detected directly by hybridization to DNA microarrays or by high-throughput sequencing. For hybridization experiments, mRNA isolated from the tissue to be analyzed is reverse transcribed into cDNA by reverse transcriptase and transferred into a labeled hybridization probe. The microarray to which this probe is hybridized contains oligonucleotides representing all possible exon sequences, which can potentially be expressed in the respective tissue of the investigated organism. The hybridization signals then indicate which genes are expressed at what copy number.

Since it is possible to immobilize millions of oligonucleotides onto a single array, whole chromosomes or even whole genomes can be displayed (tiling array) and therefore detection of transcription activity in general is possible. Hybridization experiments on these arrays showed that the number of transcripts is much higher than previously expected. Many short and long RNAs have been identified that are involved in gene regulation and enzymatic activities, but also many RNAs with so far unknown functions. The ENCODE project even determined that every sequence in the genome is transcribed at some time during an organisms life cycle in at least one direction.

Analysis of Promoter Strength and Active RNA Synthesis, Chapter 33

Protein–Nucleic Acid Interactions, Chapter 32

Alternatively, expression is determined by direct sequencing of cDNAs. High-throughput sequencing technologies allow the characterization of a complete transcriptome in a single experiment. Again, mRNA is reverse transcribed into cDNA and ligated to standard linkers to

prepare the DNA fragments for sequencing in a highly parallelized approach. The number of reads of a transcript serves as a measure for the copy number of the respective mRNA.

36.2.6 Genes and Hereditary Disease – Search for Mutations

Quantitative trait loci (QTLs) A group of variant loci within the genome that contribute to the occurrence of a certain disease.

The number of genes potentially responsible for a certain disease can be limited by the candidate gene approach (Section 36.2.4). However, the functional proof that a certain mutation is responsible for a disease phenotype still is a challenge. In addition, in many cases, a disease does not go back to a predisposition of only one gene. Instead, very often mutations of several genetic loci contribute to the characteristics of a certain disease (quantitative trait loci, QTLs).

Several approaches can help to examine the effects of mutated genetic loci. In the simplest case, when patient data gives evidence that exactly one gene is deleted and that this deletion is responsible for the deletion, the genomic region affected by deletions or translocations, respectively, can be narrowed by cytogenetic methods (FISH, fluorescent *in situ* hybridization, for example, with BAC clones as hybridization probes). FISH analysis, however, only allows mapping of very large deletions. Another classical approach to detect smaller deletions is the hybridization of the labeled chromosomal DNA to BAC arrays, containing BAC clones representing known chromosomal regions (Sections 36.2.2 and 36.2.3), which were spotted onto glass slides. If the hybridization signal is weaker than the control hybridization it points to a deletion within this chromosomal region. However, normally, several defects are involved, the affected sequences are very short (point mutations, small translocations, deletions, or amplifications), and these deviations have to be compared between many patients. To get an overview of these aberrations, several methods are available:

Methods of Fluorescence-labeled DNA Hybridization, Section 35.1

Gel Electrophoresis of DNA, Section 27.2.1

1. For the detection of single-strand conformation polymorphism (SSCP), short genomic sequences (150–300 bp) are amplified by PCR. By denaturing high-performance liquid chromatography (dHPLC), the PCR fragments differing from the wild-type sequence are detectable. This type of experiments was often performed with large patient cohorts to determine the LOH loci (loss of heterozygosity loci).
2. Exon sequencing: Exons of selected genes are PCR amplified from a patient's DNA with the aid of flanking primers. The PCR products are sequenced and the sequence is compared to the wild-type version. This allows us to look specifically for altered DNA sequences, which lead to changes in the protein sequence, potentially impairing its function.

   Both dHPLC and exon sequencing have to a large extent been replaced by high-throughput sequencing methods.

Genotyping, Section 37.2.1

3. *SNP/SNV analysis*: By hybridization of specified microarrays, single nucleotide polymorphisms (SNPs)/single nucleotide variants (SNVs) and also copy number variants (CNVs) can be detected. The hybridization of millions of SNP variants in parallel on a single array allows the determination of a whole-genome genotype of a patient, and genetic variants in patient cohorts can be detected in high-throughput (genome-wide association studies, GWAS). Alternatively, high-throughput sequencing is used to identify genetic variants (compare Section 36.2.3). Statistical methods are used to determine which markers are inherited together more often than is expected after random recombination.

   The goal of these approaches is to identify genetic markers that are relevant for future diagnostics, but also to identify drug targets for the treatment of complex diseases. One example of a worldwide research initiative is the International Cancer Genome Consortium; the ICGC project, which includes sequencing and comparison of thousands of cancer genomes. Many other diseases are under investigation to improve treatments for patients suffering from complex diseases.

36.3 Integration of Genome Maps

To get the most out of mapping approaches and functional analyses, all sorts of databases all over the world have been made available and linked. The combination of complex data allows new conclusions to be drawn from existing data.

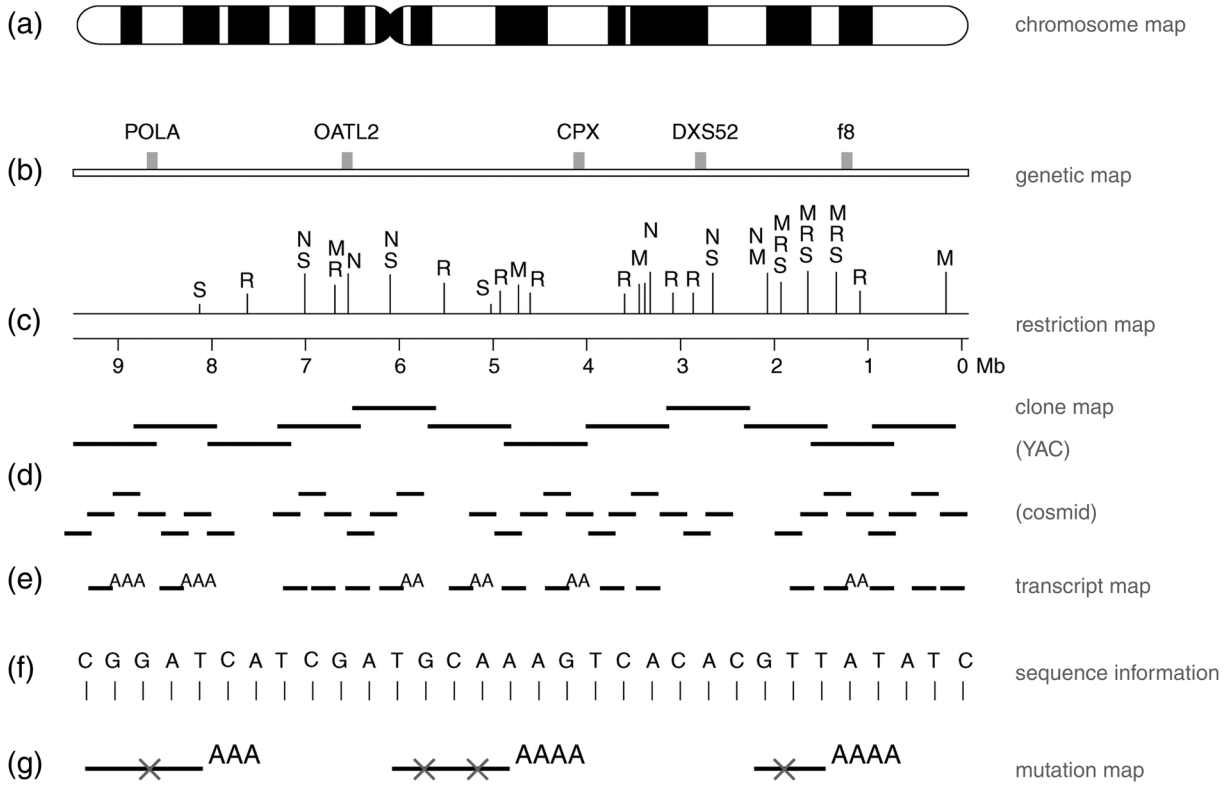


Figure 36.9 Integrated map of a genomic region. The different mapping approaches (a)–(g) are explained in the main text.

If genetic and physical maps are compared, it becomes obvious that both types of maps are collinear in most cases, as expected. In other words, the order of the markers is identical on both types of map. However, the relative distances of the genetic maps frequently do not correlate with the absolute distances of the physical maps. Figure 36.9 shows an integrated map resulting from different mapping approaches. The following list systematically address the different kinds of map in the figure:

a) The cytogenetic map is shown as pictogram of a Giemsa staining, as seen under a light microscope. The resolution of this type of map is about 5 Mb. Therefore, on this level only substantial changes to the genome can be recognized. Beside an aberrant chromosome number (e.g., human trisomy 21), these include large deletions, insertions, and translocations.
b) This genetic map portrays the correct sequential arrangement of markers along this genomic region, as confirmed by sequencing data. The recombination frequency between neighboring markers is included as well. The resolution of classical genetic maps is 0.5 cM (about 0.5 Mb) to 2 cM. The huge number of SNPs/SNVs now available allow resolution down to about 1000 bp.
c) Before high-throughput chip and sequencing technologies were available, the genetic map was often generated with the same markers as the restriction map, to be able to use equivalent reference points. This allowed a detailed correlation of both types of map. Since rarely cutting restriction enzymes only provide a resolution of several hundred kilobases, additional markers (which need not necessarily be polymorphic) were used to achieve higher resolution.
d) The clone map is based on recombinant clones. With YACs the resolution is comparable to restriction maps. With prokaryotic clones, the resolution is up to about 10 kb.
e) The classical transcription maps referred to clones that were mapped by hybridization. In addition, these maps help to isolate genes that were previously localized to certain clones. Mapping is now also possible by direct sequencing of reverse transcribed mRNA (see above).

f) Sequencing provides resolution down to the level of individual bases. Sequencing of transcripts (e) as well as the corresponding genomic region (d) allows exact mapping of mRNA (exons), but also introns and different regulatory regions, for example, 5′ and 3′ non-translated regions (5′UTR, 3′UTR) and promoter regions.
g) After sequencing of numerous individuals, different sequence variants (SNPs/SNVs) become part of an integrated map. This serves as a basis for the analysis of patients suffering from certain diseases by SSCP, direct sequencing, or microarray hybridization, to detect sequence variants that appear with a significant correlation to disease and therefore might help to identify diagnostic or even, after functional analysis, potential therapeutic targets.

Methylation Analysis with the Bisulfite Method, Section 31.2

Moreover, efforts are underway to include and present functional data in appropriate databases. Besides data about gene expression (microarray-hybridization or direct sequencing of cDNAs), for example, epigenetic data are being integrated. For example, DNA methylation profiles can be determined with the bisulfite method, in combination with hybridization of DNA microarrays or direct sequencing. The binding of DNA sequences to modified histones can be determined by chromatin-immunoprecipitation combined with microarray hybridization or sequencing (ChIP-chip, Chip-seq). In addition, the binding sites of many other proteins (e.g., transcription factors) or RNAs (e.g., micro-RNAs) can be determined. These data allow the nomination of sequences responsible for the regulation of gene-expression for verification by functional analysis. The identification of variants of respective sequences detected in patient cohorts can eventually serve as an important pre-requisite for the development of innovative therapies against some diseases.

36.4 The Human Genome

The Human Genome Project was only a milestone on the road to the understanding of human genetics. The availability of low cost, high-throughput sequencing technologies has led to an explosion in the amount of genomic data available in public databases and advances in bioinformatics have made it easy to work with resulting data. More than 10 million polymorphisms (SNPs/SNVs, insertions or deletions (indels), CNVs) – one in about 1300 bp – have been discovered so far. Low-resolution genetic maps and clone-based physical maps have been replaced by high-resolution maps at the nucleotide level. Therefore, not only many monogenic, but also polygenic diseases – diseases resulting from several genetic defects – have been mapped in great detail.

Functional and Systems Analytics, Chapter 33

The combination of genetic and physical maps greatly facilitates the study and analysis of the genetic roots of human disease. Gene expression and epigenetic data can be localized to loci thought to be associated with disease and thus enable a search for biomarkers that may help to diagnose a disease early in its course or perhaps even before a disease state has come about. Information from model organisms such as *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Mus musculus*, and their mutants (e.g., gene knock-outs), can also be consulted to narrow the search for suitable candidates. The identification of genes involved in disease can be considered in the context of functional and signaling networks, which, under the right circumstances, could lead to new therapeutic approaches to treat the disease under consideration. Thus genetic and physical mapping of genomes has become an essential part of the understanding of human genetics.

Further Reading

Adams, M.D. *et al*. (1995) Initial assessment of human gene diversity and expression patterns band upon 83 million nucleotides of cDNA sequence. *Nature*, **377**, 3–174.
Birney, E. *et al*. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
Cawley, S. *et al*. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116** (4), 499–509.
Daly, A.K. (2010) Genome-wide association studies in pharmacogenomics. *Nat. Rev. Genet.*, **11**, 241–246.

Fiegler, H. *et al*. (2003) DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer*, **36** (4), 361–374.
Hawkins, R.D. *et al*. (2010) Next-generation genomics: an integrative approach. *Nat. Rev. Genet.*, **11**, 476–486.
Hoehe, M.R., Timmermann, B., and Lehrach, H. (2003) Human inter-individual DNA sequence variation in candidate genes, drug targets, the importance of haplotypes and pharmacogenomics. *Curr. Pharm. Biotechnol.*, **4** (6), 351–378.
Hudson, T.J. *et al*. (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.
Imanishi, T. *et al*. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol., **2** (6), e162.
Krebs, J.E., Goldstein, E.S., and Kilpatrick, S.T. (2009) *Lewin's Genes X*, Jones & Bartlett, Burlington, MA.
Lander, E.S. (2011) Initial impact of the sequencing of the human genome. *Nature*, **470**, 187–197.
Maier, E. (1994) Application of robotic technology to automated sequence fingerprint analysis by oligonucleotide hybridisation. *J. Biotechnol.*, **35**, 191–203.
McPherson, J.D. *et al*. (2001) A physical map of the human genome. *Nature*, **409**, 934–941.
Nakamura, Y. *et al*. (1987) Variable number of tandem repeats. *Science*, **235**, 1616–1622.
Metzker, M.L. (2010) Sequencing technologies – the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
Ross, *et al.* (2005) The DNA sequence of the human X chromosome. *Nature*, **434**, 325–37.
Strachan, Z. and Read, A.E. (2010) *Human Molecular Genetics*, Taylor & Francis.
Venter, C. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 304–351.
White, R. *et al.* (1986) Construction of human genetic linkage maps I: progress and perspectives, in *Cold Spring Harbor Symposia on Quantitative Biology*, vol 51, Cold Spring Harbor Laboratory Press, pp. 29–38.
Yoo, S.M., Choi, J.H., and Yoo, N.C. (2009) Applications of DNA microarray in disease diagnostics. *J. Microbiol. Biotechnol.*, **19**, 635–646.
Zhang, W., Ratain, M.J., and Dolan, M.E. (2008) The HapMap resource is providing new insights into ourselves and its application to pharmacogenomics. *Bioinform. Biol. Insights*, **2**, 15–23.

DNA-Microarray Technology

37

Jörg Hoheisel

German Cancer Research Center, Head, Functional Genome Analysis & Chairman of Scientific Council, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany

Studying cellular processes at a global molecular level is prerequisite for elucidating functional mechanisms in a cell, a tissue, or an organism as a whole. DNA-microarrays are typical for this kind of research; they were the first experimental format at the level of nucleic acids that allowed various types of comprehensive analysis of many samples. The large quantities of biological information that could be gathered from these platforms were critically important for the development of biology towards a more complete rather than sketchy view of life. Only a global understanding at various molecular levels may eventually allow us to comprehend in its entirety the conversion of basic genetic information into cellular function.

Originally conceived as a procedure for mapping and sequencing genomic DNA, microarray technology developed and proliferated quickly into covering a large variety of biological analysis types. It was also instrumental for a conceptual change in biological research. First, the enormous amount of data that was generated made it essential to automate data processing, replacing classical, mostly manual analysis procedures. This kind of biology is based on numbers and statistics initially rather than observation and thereby is more quantitative. In addition, the analyses documented the high degree of dynamics that is intrinsic to biological systems. For the comprehensive nature of the information gained from microarrays, experiments could be designed that were not based on prior hypotheses. In many cases, this approach can actually be advantageous for the advancement of science, since unexpected results, which are bound to happen, are easier to accept for the lack of a preconceived opinion. In fact, the data obtained acted frequently as the nucleus for the formulation of novel theories, turning on its head the usual sequence of events.

Microarray analyses also demonstrated that an investigation – never mind how comprehensive it may be – at only one molecular level, for instance a study of the transcriptome alone, is insufficient for understanding a biological system. A vertical component is essential, linking various molecule classes and in particular also taking into account the interactions between them. Consequently, methods and tools were required for combining and merging different data sets, along with a time perspective to be considered. This led to processes that laid the basis for the development of a theoretical biology named systems biology in a process very similar to the development in physics, which split into a theoretical and an experimental branch more than a century ago.

Now, after about 25 years of use, many DNA-microarray analyses have become routine. In fact, with the advent of next-generation sequencing there is a technology around that will soon supersede microarrays in many, especially analytical, applications. However, notably, several of the current sequencing procedures actually represent array-based processes; only some more recent methods are based on other principles. For a while, DNA-microarrays will remain, but eventually they will yield to sequencing for most of the applications described below. In compensation for this loss, however, microarrays could be successful as tools for structural analyses or as a platform for molecule production and combination (really synthetic

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

biology), for instance. In addition, there are developments in fields other than the analysis of nucleic acids, such as proteomics, or for analytical processes that deal with more than one molecule class. For some of these new applications, microarrays could be the format of choice.

In this chapter, an overview is given about the current main applications of DNA-microarrays, and possible directions are discussed along which microarray technology may develop.

37.1 RNA Analyses

37.1.1 Transcriptome Analysis

For many scientists, microarray analysis is still synonymous with expression profiling, which stands for transcript analysis (Figure 37.1). This application was the first microarray format that had a strong impact on biological research. For a while, it formed the basis for studying functions encoded in the gene, which had been deciphered by the various genome-sequencing projects. This importance has clearly diminished. Apart from the fact that RNA sequencing is becoming competitive, it has also been recognized that RNA is a molecule class that may not be an optimal analyte in the first place. First, many RNAs are molecular intermediates. At the same time, mRNA especially is very dynamic in nature and regulated by very many processes simultaneously. Therefore, it is still difficult to design a model of mRNA expression, even though very large data sets on mRNA transcript profiles are available. At the same time, most RNA molecules are rather volatile and therefore less suitable for diagnostics; microRNA could actually be an exception, being found to be stable even in various body fluids. For the identification of therapeutic approaches, finally, RNA may actually be too far from the actual activity, which is mostly mediated by proteins.

Particularly in term of quantification, sequencing could be superior to microarrays. For the latter, not all experimental biases could be excluded, such as an influence of mass transport, although common minimal requirements for data quality have been defined. However, standards for data analysis, starting from basic issues like normalization, do not exist. A typical and rather common mistake, for example, is the assumption that the variation in the abundance of an RNA molecule has to be by at least a factor of two in order to be relevant. This originated from early microarray analyses of a group at Stanford University. In a concordance analysis of their data, they concluded that a change by a factor of two or more was significant for their particular experimental results. This threshold was subsequently applied to other studies, although a concordance analysis may have yielded a different value for each of them. Another flaw in most studies is the fact that, for both sequencing and microarray analysis, the focus is mostly – if not exclusively – on changes in transcript levels, and the higher they are the better. However, the degree of variation is not necessarily informative. For example, the transcript

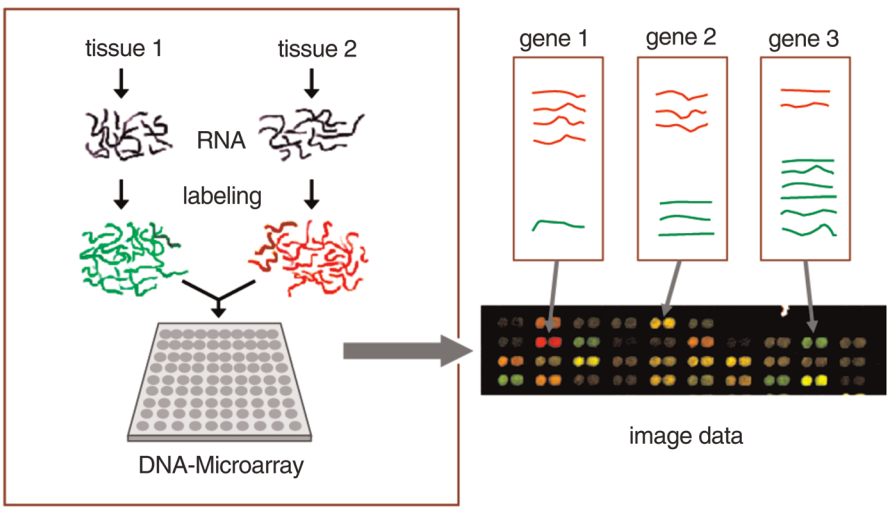


Figure 37.1 Schematic view of one format of a transcriptional profiling analysis. RNA is isolated from two tissue samples (e.g., tumor and normal tissue of a kind). One isolate is labeled with a red fluorescent dye, while green is used for the other sample. After mixing, the two samples are hybridized onto an array of DNA-fragments or oligonucleotides that represent the genes. The red and green molecules compete for the identical binding sites. In consequence, the color at each spot provides information about the ratio of the respective green and red molecules. If similar amounts are present, the resulting spots are yellow. The same information can be obtained with one dye and an incubation on two microarrays.

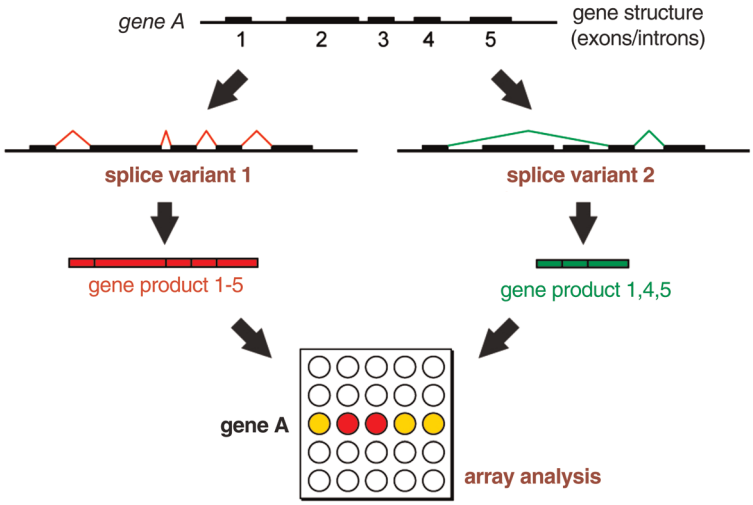


Figure 37.2 Analyzing RNA splicing. Genes of higher organisms frequently consist of several exons. During RNA processing, different exon combinations may be produced. For analysis, the RNA isolates from different samples are labeled with fluorescent dyes. After hybridization to a microarray that contains representatives of each exon, the signals indicate the differences in RNA splicing. As for transcript level profiling, two arrays and labeling with one dye would produce the same information.

level of actin is very strongly up-regulated in tumors, but is unlikely to be causatively involved in the transformation of a normal cell into a tumor cell. In addition, a lack of any change in transcript levels is nearly as important for the understanding of cell biology as a significant change, but is largely ignored.

37.1.2 RNA Splicing

Along with variations in transcript levels, differential splicing is another process responsible for regulating function. An unexpected but not really surprising result of genomic sequencing was the fact that it is not so much the mere number of genes or their sequences that is actually responsible for the differences between organisms, but the molecular modulation and interpretation of the encoded information is what matters. Information about splice variations is obtained by microarrays that contain at least one binding site for each exon of a gene (Figure 37.2). However, only exons that have been recognized by prior sequence annotation can be studied this way. Even for the relatively simple genome of the yeast *Saccharomyces cerevisiae*, several hundred exons had not been recognized right away. For the more complex genome of the fruit fly *Drosophila melanogaster*, this number was more than 2000 originally. One could argue that with the enormous and redundant amount of sequence information that is available meanwhile, there should no longer be such a problem. However, this may not be true after all. In view of the fact that the number of RNA molecules encoded in the human genome is probably in the range of 130 000 – mostly not protein-encoding genes, and many of them still not confirmed – rather than the only ca. 22 000 RNA molecules annotated originally, the number of not yet appropriately annotated exons may still be rather large.

37.1.3 RNA Structure and Functionality

The folding of RNA molecules – best known from the cloverleaf structures in which transfer-RNA (tRNA) molecules are usually visualized – can act as a strong indicator of variation in the primary RNA sequence as well as in the function of the RNA. During the early phases of microarray-based analyses, Edwin Southern and colleagues already studied the influence of RNA structure on the binding onto oligonucleotide microarrays. The array-bound oligonucleotides represented in their entirety a complementary sequence of the complete RNA molecule. Especially for RNA molecules, which either directly exhibit an activity, such as ribozymes, or act as a structural component, such as the rRNA in the ribosome, it is very likely that a change in structure goes along with a transformation of the activity or functionality. In comparison to a mere sequence analysis, the procedure has the advantage that the structural variation is tested immediately, which makes the results much more relevant.

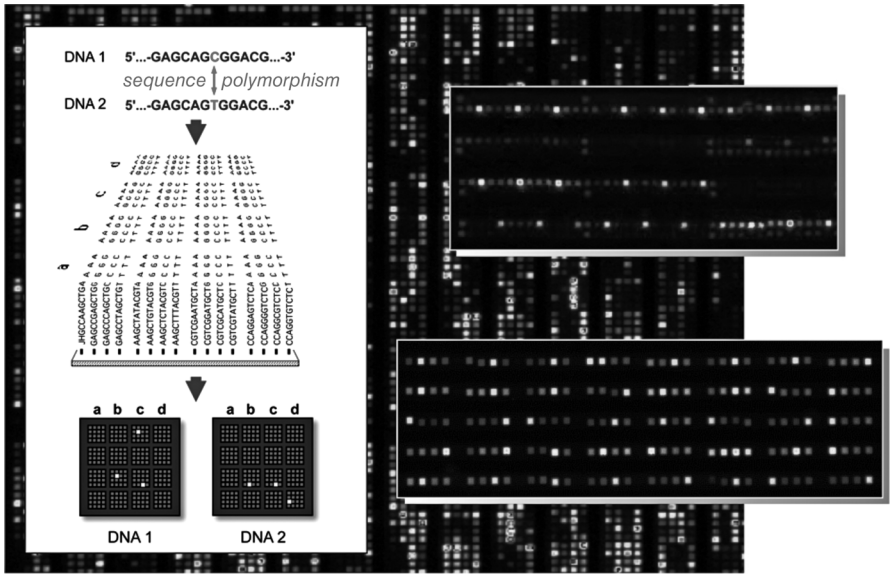


Figure 37.3 Identification of single-nucleotide polymorphisms (SNPs). Two DNA fragments, which differ in sequence by one base only, are hybridized to a microarray of oligonucleotides, which represent sequence variations. If hybridization is specific, the two DNA fragments bind to different complementary oligonucleotides. Alongside the schematic representation, examples of real data are shown. Source: adapted and modified from Hoheisel, J.D. *et al. Ann. Biol. Clin.*, 50, 827–829.

37.2 DNA Analyses

37.2.1 Genotyping

An initial objective during the development of microarray technology was the establishment of a method for a quick deciphering of DNA sequences. While this never materialized properly in the form of a hybridization-based process, it did work out with the help of enzymatic reactions (Section 37.2.3). For the identification of individual sequence variations (single nucleotide polymorphisms, SNPs), however, microarrays were and still are employed, although sequencing is bound to take over eventually.

Technically, microarray-based SNP analyses are developed to a degree that meets the requirements for clinical application. One reason is the fact that qualitative information is sufficient to call the difference. In addition, controls for an immediate internal quality assessment are readily available in the form of oligonucleotides that each contain one of the four possible nucleotides at the position in question. The actual analysis is based on the differences in duplex stability of an analyte DNA upon hybridization to probe molecules on the microarray surface that represent all four sequence variations, or 16 in case of a dimer sequence (Figure 37.3). Even more accurate results are obtained by a continuous detection of the hybridization and dissociation process between analyte DNA and probes (dynamic allele-specific hybridization). The additional information gathered from the association and dissociation curves permits an optimal discrimination between fragments with fully homologous sequences and those in which a single nucleotide (or more) is not complementary. Alternatively, an enzymatic reaction can be utilized, for example, by adding to the reaction a polymerase and dideoxynucleotides that are labeled with a base-specific fluorophore (Figure 37.4). Combining the selectivity of DNA hybridization and the specificity of a polymerase, high discrimination is achieved even for sequences that are difficult to analyze.

Genotyping is frequently used for the identification of microorganisms in diverse areas, such as the health system, food quality control, or wastewater treatment plants. In biomedicine, the creation of a high-resolution map of the human genome is basically complete; microarrays with as many as up to 10 million SNPs are used to define the position of genes that are associated with particular diseases.

37.2.2 Methylation Studies

About 4% of all cytosines in the human genome – the ones that are part of a d(CG) dinucleotide sequence, frequently called the CpG dimer in consideration of the two consecutive bases and

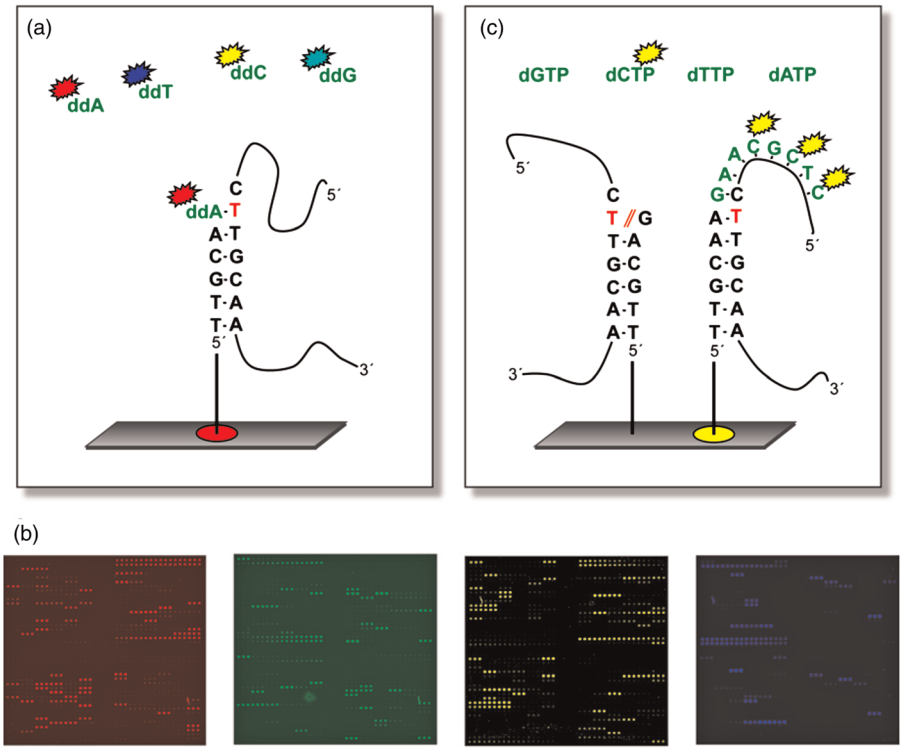


Figure 37.4 SNP analysis by means of a polymerase extension reaction. (a) In one setting, an array-bound oligonucleotide primer is used that reaches to the sequence position that is just in front of the nucleotide in question. Upon addition of the four labeled dideoxynucleotides, only the complementary nucleotide is incorporated so that the color indicates the sequence. Since the 3′-hydroxyl group is missing, only one nucleotide can be incorporated per DNA strand. In (b), typical results are shown. (c) Alternatively, the oligonucleotide primer includes the position that is being queried. Only one labeled deoxynucleotide is required but four primers; for simplicity, only two primers are shown here. Only the fully complementary primer will be extended by the polymerase, thus creating a signal at the respective array position.

the phosphate group that lies in between – exhibit variation in methylation, which is regulated by enzymes adding or removing a methyl group at either strand of the double helix. The resulting epigenetic patterns represent one of many dynamic features of DNA, which is frequently but wrongly assumed to be of a rather static nature. The binding or progression of DNA-binding proteins is influenced by the degree and position of DNA methylation, thus affecting transcription, for example. Variations in the promoter regions of tumor-associated genes play a major role in cellular transformation, for instance. Since methylation is a rather stable DNA modification, it is attractive for utilization as a biomarker. Currently, microarray analyses are still the method of choice for studying methylation at a global level but with, nevertheless, a single-base resolution (Figure 37.5). For practical reasons, only about 1.5% of all possible methylation sites are looked at in most studies. For the analysis, genomic DNA is treated with bisulfite. While methylated cytosine is not affected by the chemical reaction, the bisulfite converts unmethylated cytosine into uracil, which turns into thymine upon PCR amplification of the material. This dC into dT conversion is nothing other than a chemically induced SNP, which is subsequently studied accordingly as described above (Section 37.2.1). Alternatively to this approach, antibodies exist that bind specifically to methylated sequences. Incubation of DNA with such antibodies followed by DNA cleavage and antibody precipitation co-isolates the methylated DNA, which is then incubated on microarrays that represent many genomic regions or analyzed by DNA sequencing for identification. While microarrays are unlikely to get much beyond the analysis of a few percent of the methylation sites, although with high accuracy, sequencing will allow an analysis of all the ca. 30 million sites in a single experiment.

37.2.3 DNA Sequencing

The objective of reading DNA sequences in high-throughput was a starting point for the development of microarrays. In principle, sequencing is simply an extended genotyping. However, all possible sequence variants and all nucleotides of a DNA fragment need to be covered. One approach to achieve this was based on the hybridization of DNA to a comprehensive library of all 65 000 octamer oligonucleotides. By splitting the octamer

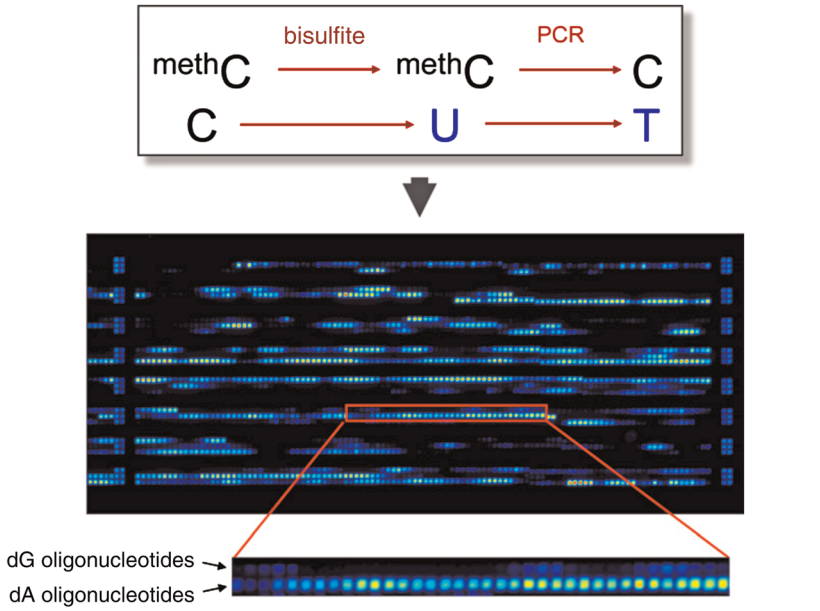


Figure 37.5 Methylation studies. Treatment with bisulfite converts unmethylated cytosine into uracil, which is turned into a thymine upon PCR-amplification of the DNA. Hybridization to a microarray that consists of the respective complement sequence in form of oligonucleotides will make the conversion, or the lack of it, apparent by the hybridization pattern produced. In the analysis shown, the DNA binds mostly to the oligonucleotides, which contain a dA at the relevant position(s) and are therefore complementary to a DNA fragment that was unmethylated originally.

sequence in two halves of four nucleotides each and placing an unspecific dinucleotide in between, thus using an array of decamer oligonucleotides, a read-length of about 2000 nucleotides would have been possible. The accuracy of this process, however, was never sufficient to compete successfully with the gel-based sequencing methods of the early 1990s.

As for genotyping, a polymerase reaction offers an alternative. The combination of the DNA binding to the arrayed oligonucleotide primers and the selectivity in extending these primers by a polymerase reaction permits high accuracy. In principle, a process that could yield long reads was already known as early as 1994, but it took more than a decade to develop the idea into a working technology. The array-bound primers are extended with nucleotides that are labeled with a base-specific dye. Only one nucleotide can be incorporated at a time as long as the fluorescence label is not cleaved off. After detection of the fluorophores at all array positions, directly indicating which bases were incorporated, the dyes are removed and another cycle of extension begins (Figure 37.6). Since this reaction takes place at very many molecules in parallel, enormous amounts of sequence data can be accumulated.

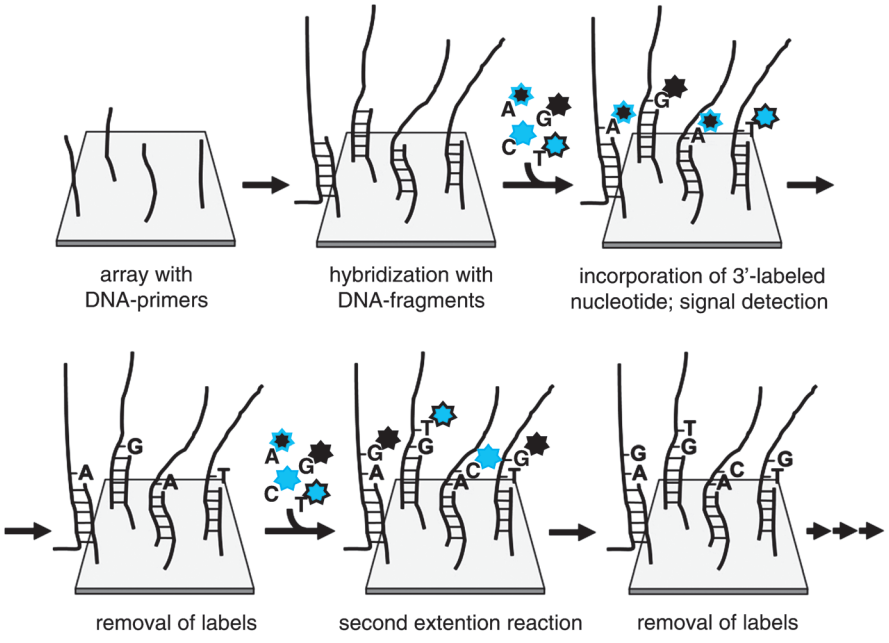


Figure 37.6 High-throughput sequencing. Surface-attached DNA primers are hybridized with genomic DNA fragments. After addition of fluorescently labeled nucleotides and a polymerase, the nucleotide is incorporated that is complementary to the respective genomic fragment. Color detection allows the base to be identified. The fluorescent dye is then cleaved off, thereby creating the 3′-end required for the incorporation of the next labeled nucleotide.

In a similar approach, the progress of the polymerase reaction is monitored by detecting the pyrophosphate, which is released upon incorporation of an (unlabeled) nucleotide triphosphate (pyrosequencing). In this process, four times as many cycles are required, however, since only one of the four bases can be added at a time. Otherwise, it would not be known whether incorporation of dATP, dGTP, dCTP, or dTTP is responsible for the release of the pyrophosphate. Pyrosequencing was the first process that yielded a next-generation sequencing device. The main obstacle in developing the process into a high-throughput method was miniaturization; the basic reaction had already been used for a while in a microtiter plate format.

Instead of a polymerase, other enzymatic reactions also allow the reading a DNA sequence. One example is DNA ligation. A mixture of short, synthetic DNA-oligonucleotides, which in combination represent all possible sequence variations of this length, and a DNA ligase are added to a primer-bound DNA template, rather than a polymerase and nucleotide triphosphates. Only one of the short oligonucleotides will then fit entirely to the sequence of the DNA template next to the primer molecule and will be incorporated by DNA ligation. Rather than base by base as in a polymerase reaction, small pieces of DNA are added consecutively. Detection is again via molecule-specific labels.

37.2.4 Comparative Genomic Hybridization (CGH)

Analyzing the copy number of particular genomic regions, named comparative genomic hybridization (CGH), is an important diagnostic procedure for the detection of genetic aberrations. Usually, there are two copies (alleles) of a genome in each human cell, with the exception of most of the genetic content of the X-chromosome in a male person. However, the local copy number can vary largely and differ between people or between healthy and diseased tissues. This is analyzed by hybridizing labeled genomic DNA onto a representation of the genome or parts thereof. A signal intensity that is stronger or weaker than that obtained with a reference DNA indicates regions in which amplifications or deletions have taken place. Especially for tumors, the connection between such variations and disease has been documented and is being used as a diagnostic tool. However, such variations also occur in healthy tissues, acting as a normal regulation process. Originally, the analyses were performed on metaphase chromosome spreads. A microarray version, however, simplified the handling and improved resolution. Initially, the array-bound fragments had a length of about one million base pairs each; nowadays, microarrays of short oligonucleotides are used for a high-resolution scanning down to few kilobase pairs. Analysis by DNA sequencing pushes the resolution to its ultimate limit – one base pair – and produces more quantitative results, since the frequency of a regional sequence can actually be counted.

37.2.5 Protein–DNA Interactions

A central component in the regulation of transcription is the binding of transcription factors to the promoter regions of genes. However, many other interactions of proteins and DNA are also important for the functioning of a cell. The position of this kind of interaction is frequently determined by chromatin immunoprecipitation and a subsequent analysis of the isolated DNA by genome-representing microarrays (ChIP-on-chip; Figure 37.7) or DNA sequencing. To this end, all DNA-bound proteins are chemically crosslinked to the DNA. An antibody is used to isolate (precipitate) the protein in question. Because of the crosslink, the protein-bound DNA fragment is co-precipitated. After removal of the protein by an enzymatic digestion, the DNA can be analyzed. For this kind of analysis, both the specificity of the protein–DNA interaction and the selectivity of the antibody are critical. The procedure was instrumental for obtaining functional information about the cellular regulation of transcription in yeast, for example. Even in stationary cells, the entire protein machinery needed for transcription was found to be present, but in an inactive state. RNA polymerase II sits in the promoter regions of several hundred genes that are critically important for a quick response of the yeast cells to a change in the environmental conditions, when nutrition becomes available.

However, microarray analyses also provide quantitative information about the specificity and intensity of protein–DNA interactions. In fact, this aspect will be an important application of microarrays in future, while the mere determination of binding sites will be done by sequencing.

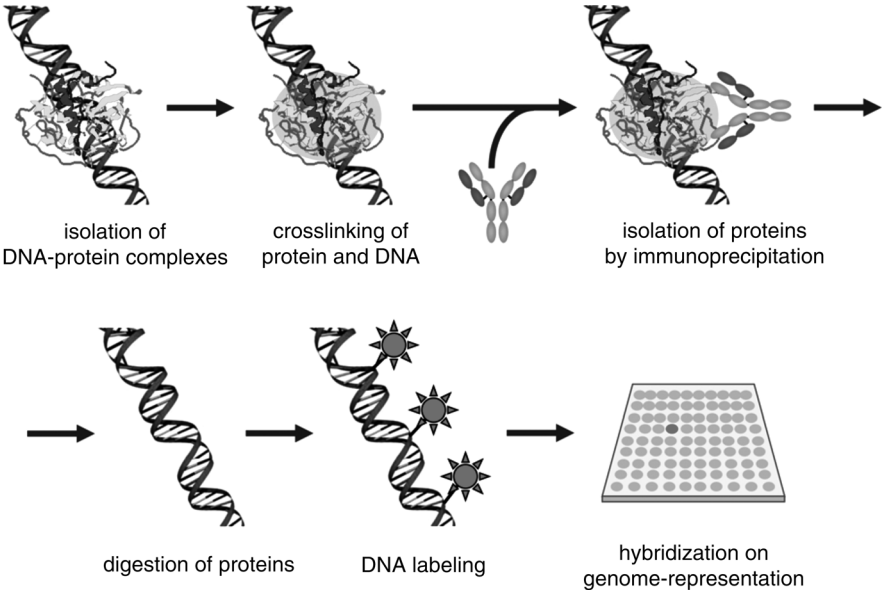


Figure 37.7 ChIP-on-chip analysis. Protein–DNA complexes are crosslinked in a chemical reaction. A particular protein is then isolated by means of an antibody. The co-isolated DNA fragment is released by digesting the protein enzymatically. After labeling with a fluorescent dye and hybridization to an array that represents the genome or a portion thereof, the location of the DNA and thus the binding position of the protein is identified.

Most sequence-specific proteins actually bind to double-stranded DNA. Therefore, processes had to be established to produce double-strand DNA on microarray surfaces. Apart from spotting pre-produced molecules (e.g., PCR products) it is also possible to create double-strand DNA from single-stranded oligonucleotides *in situ*. One option is a synthesis of long oligonucleotides, which are self-complementary in sequence and fold back on themselves, forming hairpin structures. Alternatively, a short terminal self-complementary sequence could act as a primer site for a polymerase reaction. Such an approach requires the attachment of the oligonucleotide to the solid support via its 5′-end, which in turn requires a chemistry for DNA synthesis that is identical to the natural direction of enzymatic DNA synthesis but reverse to the 3′-5′ direction of standard chemical synthesis. To analyze the binding behavior of transcription factors, for example, microarrays have been produced that contain all possible 10 bp double-strand sequences.

37.3 Molecule Synthesis

37.3.1 DNA Synthesis

Principles of the Synthesis of Oligonucleotides, Section 27.6.1

Oligonucleotide synthesis is very much automated nowadays, based on a combination of solid-support synthesis protocols and phosphoramidite chemistry. For biomedical applications, there have been two diametric tendencies during recent years. For some applications, such as the use of oligonucleotides as therapeutic agents, gram to kilogram amounts of relatively few molecules are needed. In contrast, a few picomoles or even femtomoles of an oligonucleotide are sufficient in many areas of molecular biology. By design, oligonucleotide microarrays are an ideal tool for the synthesis of many oligonucleotides with different sequences in small quantities. Systems have been established that allow the synthesis of DNA stretches, which put together represent complete genes or even entire microbial genomes. By programmable *in situ* synthesis, controlled by light induction, for example, any sequence can be produced. Since the yield of chemical synthesis on microarrays is quantitative, relatively long fragments are within reach. In addition, both synthesis directions are made possible. Thus, either the 3′-ends or the 5′-termini of the products are well defined. In a 5′-3′ synthesis, the means exist to block chemically all truncated molecule derivatives, which are short of the desired product, so that only full-length oligonucleotides act as substrates in subsequent reactions, such as a polymerase extension. For double-strand formation, the oligonucleotides are cleaved off the microarray surface after synthesis; the eluted material finds its complementary sequence in solution or by hybridization to another microarray, on which the complementary oligonucleotides have been synthesized.

37.3.2 RNA Production

Functional studies are critical for understanding the information that is encoded in a genomic sequence. Knockout or knockdown experiments are essential to such ends. Technology in this field has advanced greatly in recent years with the advent of methods such as the CRISPR-Cas9 system to target basically any genomic region. Another successful technology is based on the use of short-hairpin RNA (shRNA) constructs, which are delivered to cells via a lentiviral system and act as inhibitory RNA (RNAi) molecules, knocking down the expression of their target genes. Beside such artificial systems, more and more natural RNAi molecules are being found, which regulate transcription of a gene or a group of genes. Both natural and artificially designed inhibitory transcripts exhibit strong differences with regard to their effectiveness; inhibition is basically never complete. In addition, it is well established that even a complete deletion of a particular gene may not necessarily result in phenotypically identifiable changes, because of compensation processes at the molecular level. Therefore, a mixture of RNAi activities may be needed to produce a phenotypical result.

To work out the effect of naturally occurring RNAi molecules individually and in combination, for the identification of the best gene-inhibiting molecule, or to establish the optimal RNAi mixture for obtaining a particular phenotype, microarrays provide a rather simple process for the production of inhibitory transcripts. This could be cheaper and more flexible than producing a comprehensive library, of which only small portions would be used. Concerning the production of RNAi pools, complexity is limited to no more than a few hundred molecules; otherwise, the viscosity of the resulting solution would be too high for the solution to be handled.

For RNA production, DNA-oligonucleotides are synthesized on the microarray. Next to the complement to an RNAi sequence, they contain the promoter sequence of an RNA-polymerase, frequently T7 RNA-polymerase. The promoter is put in place as a complete fragment in a single chemical synthesis step. The parallel synthesis of the individual molecules is then performed on this common promoter fragment. As a last step, again an entire fragment is added, this time a T7 terminator adapter. By on-chip PCR, all oligonucleotides are made double-stranded. These molecules are used as templates for RNA production. Either they are first released from the microarray, eluted, and used as templates in solution or the templates remain bound to the microarray and only the enzymatically synthesized RNA is eluted. Because of the enzymatic production process with T7 RNA-polymerase, amounts of RNA can be produced that are relatively large with respect to applications in molecular biology. Since the RNA production has similar yields for each individual RNA molecule – being short and with the transcription initiated on the same promoter – the complexity of a mixture and their ratio is defined by the frequency with which the respective oligonucleotides are present on the array surface.

37.3.3 On-Chip Protein Expression

Quite a few proteins do not express well in *Escherichia coli* or other cellular systems. In addition, the entire process of cloning, transfer, expression, and purification is rather complex and elaborate. As an alternative, several different cell-free systems have been established for protein expression. Their combination with an *in situ* synthesis on template DNAs that are arranged on microarrays offers an avenue to produce very many proteins in parallel and make them accessible to subsequent analyses, such as interaction studies. Already during the synthesis process, each protein is modified, for example, by adding a terminal histidine tag, so that it binds to the microarray surface immediately next to the DNA template at the location of its synthesis. Thereby, complicated isolation and purification processes are avoided. Any modification in the DNA sequence will be readily translated into the respective variation of the protein.

Even an actual protein array printing is made possible. For this, the initial microarray with the PCR products is covered with a second, empty microarray. By virtue of appropriate surface chemistries, the *in vitro* synthesized proteins do not bind to the DNA-microarray but to the surface of the second microarray, which thereby carries a protein copy of the initial DNA pattern. Since the DNA-microarray is not affected in any way during the protein expression, the process of protein production can be repeated several times over, producing many protein arrays from a single DNA-microarray. The technology of producing protein microarray by *in situ* protein expression (Figure 37.8) has been around for more than a decade now, and several

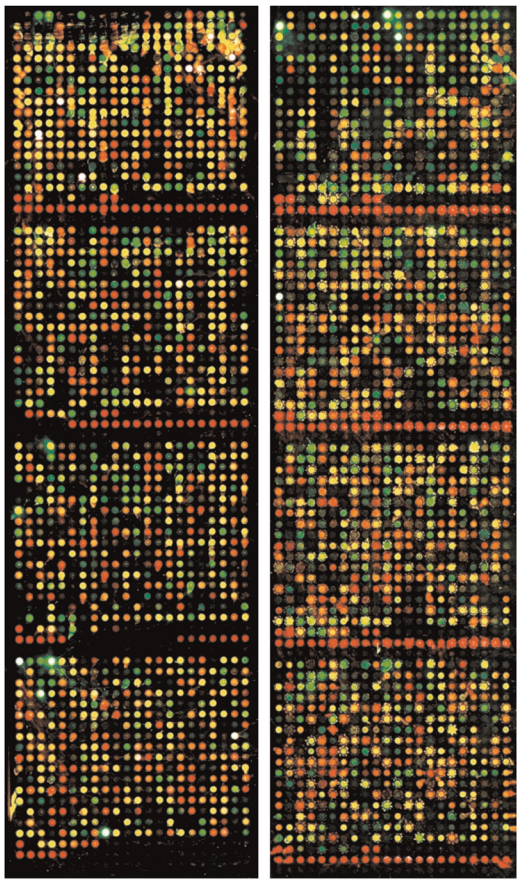


Figure 37.8 *In situ* protein expression. PCR products were placed onto two microarrays, which represented about 1600 human genes each. After parallel *in situ* transcription and translation on each individual PCR fragment, the newly synthesized proteins were detected with two red and green labeled antibodies that bind specifically to the amino- or carboxyl-end, respectively. From the color at the individual microarray positions it could be concluded that about 70% of the proteins were present in full length. Meanwhile, technology has advanced and the rate of full-length protein is now usually in the range of 95%.

technical modifications exist, but several quality measures are still to be defined and established before the technology's true potential can be utilized.

37.4 Other Approaches

37.4.1 Barcode Identification

For functional studies, more and more artificial DNA molecules are introduced into cells in order to determine the changes that occur upon addition or replacement of a particular genomic segment. Because of the complexity of regulative processes, but simultaneously driven by new capabilities made possible by new technical developments, these studies are based on the addition of different DNA constructs to individual cells, and this in very many combinations. To keep track of this, barcode sequences are frequently utilized. Barcodes are synthetic DNA sequences that do not occur in any known organism. With common primer binding sites to the left and right, they can be isolated from the genomic DNA by PCR. The first genome-wide use of this approach was a library of knockout yeast mutants. In each mutant, a particular gene was deleted; instead, two barcode sequences were introduced that were specific for the respective mutant. The library was incubated under various growth conditions, for example, and the barcode sequences were used as an indicator of whether a particular mutant cell was growing well or badly in comparison to all the others.

Another, more recent example of such an approach is shRNA libraries for genome-wide gene knockdown (Figure 37.9). Each shRNA sequence is accompanied by a specific barcode sequence of 60 bp, which is again amplifiable with a common primer pair. The shRNA constructs are transduced into host cells with the help of lentiviruses. Each construct integrates into the genome of its host cell, from which the RNAi is then transcribed at a very high and constant level. Currently, up to 55 000 different shRNA constructs can be used

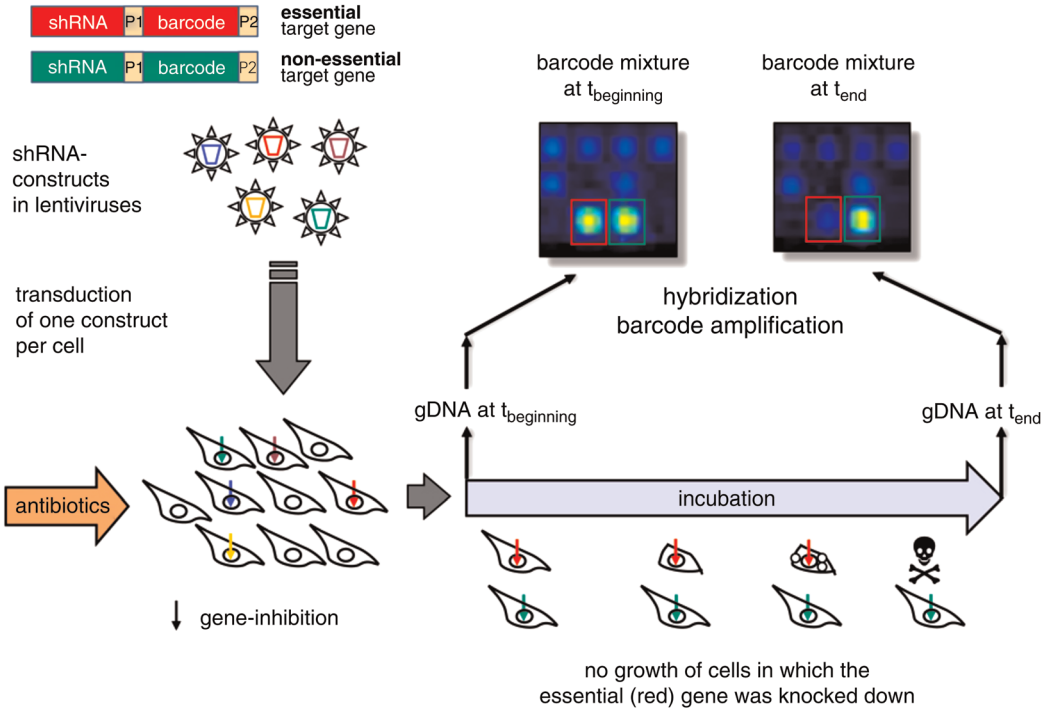


Figure 37.9 Schematic representation of a genome-wide gene inhibition analysis. A mixture of shRNA constructs is transduced into cells by means of lentiviruses. The red construct represents a shRNA, whose target gene is essential for cell survival. The green construct knocks down a gene that is not essential. All cells that do not contain any of the constructs are killed by addition of antibiotics. The remaining cells are then incubated at selective conditions. At the beginning and end of the incubation period, the genomic DNA of the cells is isolated, from which the barcode sequences are PCR-amplified with a common primer pair. Hybridization to an array of the barcode sequences shows that the relative number of "green" barcode sequences and thus "green" cells remained the same, while the "red" cells have disappeared, as the target gene was essential for survival.

simultaneously (one construct per cell). Growing the cell mixture under any sort of selection pressure will increase the number of cells for which the particular RNAi-triggered knockdown is advantageous and decrease the cell number if the knockdown is a disadvantage. The frequency of all barcodes – which is equivalent to the number of cells containing the respective construct – is determined at the beginning of the experiment and at its end. Any significant variation between the two measurements indicates an effect of knocking down a particular target gene. From the barcode sequence, it is immediately known which shRNA, and thus which gene, is involved. Until recently, this read-out was performed by hybridization of the barcode PCR-products to microarrays. The signal intensity acted as a measure of frequency. Nowadays, next-generation sequencing permits a better accuracy, since the frequency of each barcode is counted.

37.4.2 A Universal Microarray Platform

Most microarray platforms are designed for one particular type of assay. This means that a specific microarray has to be produced for each analysis form and each organism. In addition, while the physical separation of reactions is useful for analyzing complex processes, the array surface is an inhibitory factor for many assay reactions; they work better in solution than on a solid support. To circumvent all this, a common basic microarray format could be created that is used only for the data read-out, isolating, and separating a mixture of molecules. This molecular mixture is responsible for the actual assay, which is performed in solution prior to the array-based analysis. To such ends, a special kind of "barcode array" is required, usually called a zip-code microarray. Zip-codes are oligonucleotides that are unique in sequence and do not cross-hybridize. In addition, they exhibit very similar thermodynamic parameters so that hybridization works equally well on all of them. The actual assay is done with a mixture of free oligonucleotides; in addition to the sequences needed for the assay, they contain tag-sequences

that are complementary to the zip-codes. After the assay is performed in solution, the material is incubated on the zip-code microarray, upon which the assay molecules bind to separate array position and can thus be read-out individually. To avoid any cross-hybridization between assay molecules and zip-codes, the use of enantiomeric L-DNA – the mirror-image form of natural DNA, forming a left-turning double helix – has been suggested for the zip-code oligonucleotides and the tag sequences, while the assay oligonucleotides consist of natural D-DNA. Since there is no interaction between L-DNA and D-DNA, the two oligonucleotide groups could only hybridize to each other via the complementary L-DNA sequences.

37.5 New Avenues

37.5.1 Structural Analyses

An analysis of the structural variations of DNA is still an underdeveloped and underestimated field, although their role in regulative processes, and thus their functional consequences, is likely to be enormous. Already the helical angle in a DNA double helix varies between 30° and 40°. The actual structure is dependent on the sequence, and also on environmental parameters, and is of central importance to the activity of many DNA-binding proteins. In addition, DNA conformations exist that are very different from the typical right-helical image we all have in mind when thinking about DNA. Sequences of alternating purines and pyrimidines, for example, and particularly regions that consist of repetitions of the dinucleotide d(CG) – even more so when they are methylated – can flip over into a left-helical Z-DNA conformation under physiological conditions. Intriguingly, the number, position, and length of CpG stretches in promoter regions and concurrently their methylation state influence the DNA structure, and thereby modulate in a reversible manner the binding behavior of proteins that are important for transcription. To date, no direct functional involvement of Z-DNA sequences has been documented, although Z-DNA binding proteins do exist and Z-DNA does occur in living cells. However, the CpG sequences may not actually need to be in the Z-configuration to exhibit their function but may only alter the overall DNA topology of longer DNA stretches to make their influence matter.

DNA topology is a feature that is only partly sequence dependent. Sequence may actually act more as a means for the fine-tuning of topologically induced variation, defining where and under what conditions structures such as cruciform DNA, bend DNA, or single-strand stretches occur. Basically any sequence could switch into Z-conformation, for example, if enough topological stress is provided. Correspondingly, a transition of Z-DNA into a B-conformation in even a small DNA fragment will change the superhelical status of long DNA stretches. Quite a few genes are known whose transcription is directly dependent on the degree of DNA superhelicity. All this can be studied in much detail on microarrays. Attachment of both ends of a DNA fragment to the array surface allows the enzymatic introduction or removal of any number of turns, in order to create DNA pieces with structural features at a higher level than the double helix, so called supercoils. It is known that DNA has the capacity to store information not only in its sequence but also in its three-dimensional structure. Array-based studies could therefore lead to new insights into DNA-based regulation processes.

37.5.2 Beyond Nucleic Acids

Many technical aspects that were developed for the microarray analysis of nucleic acids are being adapted to studies on other molecule classes or biological systems, such as proteins, tissue sections, or even living cells. Particularly in the area of proteome analysis, microarray-based analysis forms have advanced enormously. Proteome-based diagnostics is bound to be more informative than studying nucleic acids, since proteins are usually the direct effectors in cellular processes. In addition, about 80% of the proteins assemble in multiprotein complexes, and microarrays offer a platform for analyzing the very many interactions in such complexes in a quantitative manner. Processes exist to express all human proteins – assuming one gene encodes one protein – on a single array (Section 37.3.3), and this number is bound to increase in order to also cover protein derivatives. Antibody microarrays are used for an analysis of

expression variations as well as studies on structural changes or post-translational modifications. All this could be done simultaneously. What is missing most for protein studies is a comprehensive coverage with antibodies. Although nearly 2.95 million antibodies targeting 19154 proteins are listed in Antibodypedia (status in November 2017), a database of available antibodies, coverage is far from sufficient. Many binders lack specificity or exhibit inadequate affinities, others bind to denatured proteins only or do not recognize modifications, such as polysaccharides. In addition, many antibodies, and this also includes monoclonal ones, are not a resource that is reliable enough for long-term reproducibility of their characteristics. Recombinant systems may be a better option in the long term – also because antibody identity is confirmed by sequence.

The capability of revisiting microarray spots with high accuracy, thereby introducing the option of performing different types of analysis subsequently to each other – for example, analyzing the bound material by mass spectrometry after an optical measurement of signal intensities – offers an analysis of more and more molecule classes in combination. Expansion of this development towards ever more complex technical platforms that generate data from ever more complex biological systems, *in vitro* or *in vivo*, is a forthcoming direction of microarray technology, while many of the merely DNA-based analyses will soon be replaced by sequencing. Integration of information is already an essential part of data analysis and interpretation, and this will not be limited to *in silico* studies in future but be validated and utilized on novel, integrative experimental platforms.

Further Reading

Early, Conceptual Publications

Cantor, C.R., Mirzabekov, A., and Southern, E. (1992) Report on the sequencing by hybridisation workshop. *Genomics*, **13**, 1378–1383.

Drmanac, R., Labat, I., Brukner, I., and Crkvenjakov, R. (1989) Sequencing of megabase plus DNA by hybridisation: theory of the method. *Genomics*, **4**, 114–128.

Gains, W. and Smith G. (1988) A novel method for nucleic acid sequence determination. *J. Theor. Biol.*, **135**, 303–307.

Khrapko, K., Lysov, Y., Khorlyn, A., Shick, V., Florentiev, V., and Mirzabekov, A. (1989) An oligonucleotide hybridization approach to DNA sequencing. *FEBS Lett.*, **256**, 118–122.

Maskos, U. and Southern, E.M. (1992) Oligonucleotide hybridisations on glass supports: a novel linker for oligonucleotide synthesis and hybridisation properties of oligonucleotides synthesised in situ. *Nucleic Acids Res.*, **20**, 1679–1684.

Poustka, A. *et al*. (1986) Molecular approaches to mammalian genetics. *Cold Spring Harbor Symp. Quant. Biol.*, **51**, 131–139.

Other Relevant Publications

Berger, M.F. and Bulyk, M.L. (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protocol*, **4**, 393–411.

Betzen, C., Alhamdani, M.S.S., Lueong, S., Schröder, C., Stang, A., and Hoheisel, J.D. (2015) Clinical proteomics; promises, challenges and limitations of affinity arrays. *Proteomics Clin. Appl.*, **9** (3-4), 342–347.

Brazma, A. *et al*. (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.

Hoheisel, J.D. (2006) Microarray technology: beyond transcript profiling and genotype analysis. *Nat. Rev. Genet.*, **7**, 200–210.

Moffat, J. *et al*. (2006) A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell*, **124**, 1283–1298.

Ramachandran, N. *et al.* (2004) Self-assembling protein microarrays. *Science*, **305**, 86–90.

Schena, M., Shalon, D., Davis, R.W., and Brown, E.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Tian, J. *et al*. (2004) Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*, **432**, 1050–1054.

The Use of Oligonucleotides as Tools in Cell Biology

38

Erik Wade and Jens Kurreck

Berlin University of Technology, Institute of Biotechnology, TIB 4/3-2, Department of Applied Biochemistry, Gustav-Meyer-Allee 25, 13355 Berlin, Germany

Since the beginning of the post-genomic era, which began with the completion of the sequencing of the human genome, the greatest challenge to applying this wealth of knowledge is the analysis of the function of the approximately 20 000–25 000 human genes and their relevance to human diseases. One common way to investigate the role of a gene is to block its function and analyze the resulting loss of function phenotype. One of the best ways to do this is by using oligonucleotides that specifically bind to the mRNA and inhibit gene expression. Since they are complementary to the mRNA, which is, by definition, the sense-strand, these molecules are referred to as antisense oligonucleotides. Expressed in pharmacological terms, the target sequence in the mRNA is the specific "receptor" for the antisense oligonucleotide. After hybridization to a target RNA, the antisense oligonucleotides work in two ways: they inhibit translation by sterically blocking the ribosome and by triggering the degradation of the bound RNA by cellular RNases.

In the early 1980s it was discovered that not only proteins but also oligoribonucleotides (short RNA fragments) can have catalytic activity; at the time this was quite surprising. These RNA enzymes are called ribozymes and can also be used to specifically inhibit the expression of a gene. Similar to antisense oligonucleotides, they bind to an mRNA through specific base pairing; they have the additional ability to completely inactivate the target mRNA by cleaving it into two or more parts, without the aid of proteins.

A more recent and very promising discovery is the phenomenon of RNA interference (RNAi). This involves short double-stranded RNA molecules, called small interfering RNAs (siRNAs) that cleave and degrade an mRNA with the aid of cellular enzymes. This particularly efficient mechanism of post-transcriptional gene silencing can be used, like the previously described ones, to specifically inactivate a gene. Small double-stranded RNA molecules are, however, not only artificial tools useful for research purposes and for therapies, they also play a very important role as natural regulators of gene expression. These molecules, denoted as microRNAs (miRNAs), are involved in many normal physiological processes and pathological events and have increasingly become the focus of recent research.

The use of oligonucleotides to inhibit gene expression promises levels of specificity unknown for small molecules. Most current medicines are small molecule compounds, which bind to proteins and inhibit a catalytic center, for example, or block a receptor. Unspecific, or at least unintended, binding to other proteins can lead to side effects. Antisense oligonucleotides, ribozymes, and siRNAs can be so conceived that they are significantly more specific for the target molecule. There is a very high statistical probability that a sequence of 15–17 base pairs only appears once in the human genome. An antisense molecule of this length hybridizes with this single type of mRNA and inhibits its expression.

The described techniques can be categorized as anti-mRNA approaches, since the oligonucleotides bind to a target mRNA through complementary base-pairing in each case. The oligonucleotides, or the ribonucleases (RNases) they recruit, work like molecular scissors and inhibit the expression of genes through the degradation of the mRNA. Oligonucleotides possess

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

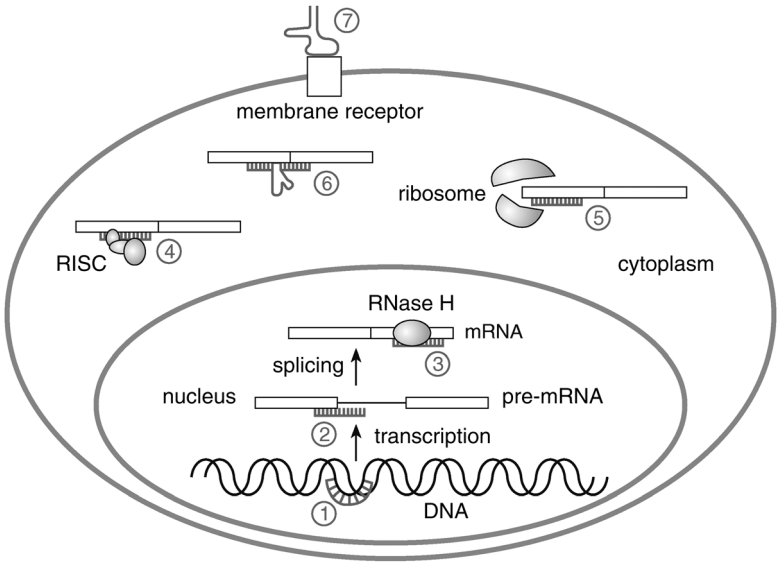


Figure 38.1 Mechanism and location of oligonucleotides used as tools in cell biology: (1) Triplex forming oligonucleotides bind to DNA double strands and block transcription. (2) Newly transcribed pre-mRNA contains introns (single lines) between the exons (rectangles) that need to be spliced out. Antisense oligonucleotides modulate splicing. (3) In the nucleus, antisense oligonucleotides induce the cleavage of pre-mRNA or spliced mRNA by RNase H. (4) In the cytoplasm, short, double-stranded RNA molecules (siRNAs) trigger the phenomenon of RNA interference (RNAi). A target RNA is cleaved by the RNA-induced silencing complex (RISC). (5) In addition, antisense oligonucleotides inhibit translation by blocking initiation of the ribosome (shown here) or elongation. (6) Ribozymes bind to the target RNA through Watson–Crick base pairing and cleave them, without the aid of proteins. (7) Aptamers bind target structures with high affinity. They are often directed against extracellular molecules (signaling molecules, membrane proteins), but can also be used intracellularly.

another interesting property, which was ignored for a long time: They form complex secondary and tertiary structures and can interact through these structures, similar to antibodies, with other macromolecules in cells, such as proteins, carbohydrates, lipids, and nucleic acids. In the early 1990s, a combinatorial approach was developed to identify an oligonucleotide with a structure that has a high affinity for a target, for example, a special protein. In this procedure, a very large number of oligonucleotides with different sequences are created and those with the highest affinity to the target are selected in each round, which is repeated until an oligonucleotide with the desired properties is found. Such an oligonucleotide is called an aptamer (from Latin *aptus,* fitting).

As summarized in Figure 38.1, oligonucleotides can be used in different ways to investigate gene function in functional genomics. They are not only of interest to basic researchers, they are also used in drug discovery to determine if a new target for a drug, for example, a new cancer treatment, plays, in fact, a central role in the course of the disease (target validation). Finally. oligonucleotides can be used as therapeutics.

38.1 Antisense Oligonucleotides

Rous sarcoma virus is an avian retrovirus that can trigger the growth of a tumor.

In 1978 the American scientists Paul Zamecnik and Maria Stevenson first described the principle that the expression of a gene can be specifically inhibited by a relatively short, synthetic oligodeoxyribonucleotide. They synthesized a 13 nucleotide long antisense oligodeoxyribonucleotide, which was complementary to a region of the Rous sarcoma virus (RSV). Addition of the oligodeoxyribonucleotide to the cell culture slowed the growth of the virus. Notably, the authors already postulated that the principle could be used therapeutically, that the RNA or DNA sequence of a gene could serve as a receptor for a pharmacologically active oligonucleotide and that the interaction should be specific.

Antisense oligonucleotides hybridize with target sequences in the single-stranded mRNA, which are complementary to their own sequence. They function by different mechanisms, of which the three most important will be described in more detail:

- induction of a RNA degrading enzyme, ribonuclease H (RNase H),
- inhibition of translation,
- changes in RNA splicing.

38.1.1 Mechanisms of Antisense Oligonucleotides

Induction of RNase H The most important mechanism that leads to degradation of a target mRNA is the induction of the cellular endonuclease RNase H. It is expressed in almost all cells

and plays an important role in DNA replication. The relevant characteristics of RNase H are its ability to recognize DNA-RNA hybrids and to degrade the RNA portion of such hybrids. Precisely this mechanism is used in antisense approaches (Figure 38.2): an antisense oligodeoxyribonucleotide hybridizes with its complementary mRNA target sequence and a DNA-RNA hybrid is formed. RNase H binds to this hybrid and destroys the RNA by endolytic cleavage.

In vitro experiments proved that the oligonucleotide must be at least four nucleotides long for the resulting hybrid to be recognized and degraded by RNase H. Corresponding to its physiological function in DNA replication, the majority of the cellular RNase H activity is found in the nucleus. What part of the mRNA an oligonucleotide binds to is not important in triggering RNase H degradation; however, experiments have shown that the efficiency of this induction can vary a great deal. The current idea is that mRNA forms complex secondary and tertiary structures and binds proteins, so that not all segments of the RNA are equally accessible for hybridization with oligonucleotides. Since the endonuclease activity of the RNase H is only activated when both the oligonucleotide and the enzyme bind to the target RNA, it is understandable why certain RNA sequences are more sensitive to oligonucleotide-induced degradation than others.

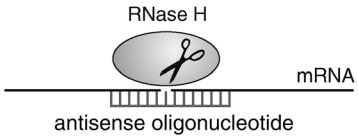


Figure 38.2 Induction of RNase H by antisense oligonucleotides. When an antisense deoxyribonucleotide binds to a complementary mRNA, a DNA-RNA hybrid is formed, which is recognized by ribonuclease H. This results in the cleavage and degradation of the RNA component of the heteroduplex. The oligonucleotide can dissociate and induce the degradation of another mRNA molecule.

Inhibition of Translation The position of the target sequence in the mRNA is very important for another antisense mechanism, the inhibition of translation by the blockade of the ribosome by antisense oligonucleotides. If translation is to be inhibited by blocking the binding of the ribosomes to mRNA, the sequence of the antisense oligonucleotides must be so chosen that it hybridizes in the 5′ region of the mRNA. If the antisense oligonucleotide binds to the translated portion of the mRNA instead, movement of the ribosome along the mRNA is blocked, thereby inhibiting elongation. It has been shown, however, that a ribosome that is reading an mRNA can, under certain circumstances, displace bound antisense molecules, so that inhibition of translation initiation is usually the more efficient way to inhibit gene expression.

Changes in RNA Splicing The transformation of the information contained in a DNA sequence into a corresponding protein sequence begins with the messenger RNA synthesis by a type II RNA polymerase. The resulting primary transcript is a pre-mRNA, which contains non-coding introns, in addition to the coding exons. The maturation of the mRNA requires splicing and takes place on specific, well-conserved sequences, the splice acceptor and splice donor sites. If an antisense oligonucleotide binds to one of these sequences, the splice site is no longer recognized by proteins and RNAs of the spliceosome and the maturation of the mRNA is inhibited.

Through the blockade of a splice site it is not only possible to inhibit the expression of a gene, it is also possible to deliberately change splicing and thereby induce therapeutic effects. For example, the β-globin mRNA is incorrectly spliced in one form of β thalassemia, resulting in a lack of a functional protein capable of transporting oxygen (Figure 38.3). If the incorrect splice site is blocked, it is skipped and the mRNA is correctly spliced such that functional β-globin is produced. Clinical trials using oligonucleotides to mask a splice site, causing the spliceosome to splice out the defective exon, are underway for Duchenne muscular dystrophy. The oligonucleotide must be composed of modified components that do not induce RNase H activity since the mRNA should not be degraded by RNase H in this case (see Section 38.1.3).

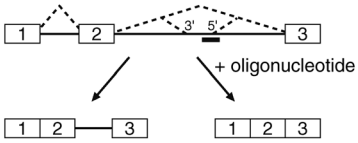


Figure 38.3 Alternative splicing of the human β-globin pre-mRNA. The exons are shown as rectangles, the introns as solid lines. The dotted lines show the various splice variants. Additional splice sites are present in β-globin as a result of mutations in β thalassemia patients (labeled 3′ and 5′). The resulting mRNA contains a piece of the intron between exons 2 and 3 (left), so that a functional protein cannot be synthesized. An antisense oligonucleotide directed against the extra 5′ splice site (thick, short line) leads to the correct splicing of the pre-mRNA (right).

38.1.2 Triplex-Forming Oligonucleotides

An interesting alternative to the previously described antisense oligonucleotides, which work at the level of mRNA, are oligonucleotides that bind to a DNA double strand, thus forming triple helices. This is referred to as an anti-gene strategy, since the triplex forming oligonucleotides (TFOs) work directly on gene expression by blocking transcription, rather than translation.

TFOs bind to DNA strands with a long sequence of purines (adenine and guanine). While the two complementary strands of the DNA duplex are bound together by Watson–Crick hydrogen bonding, additional bonds to the additional oligonucleotide are formed via Hoogsteen base pairs (Figure 38.4). In the first example of the figure, a cytosine forms hydrogen bonds with a G/C base pair. In the second example a thymine interacts with an A/T base pair. However, Hoogsteen base pairing is not restricted to the two types shown but instead can take place between other nucleotides.

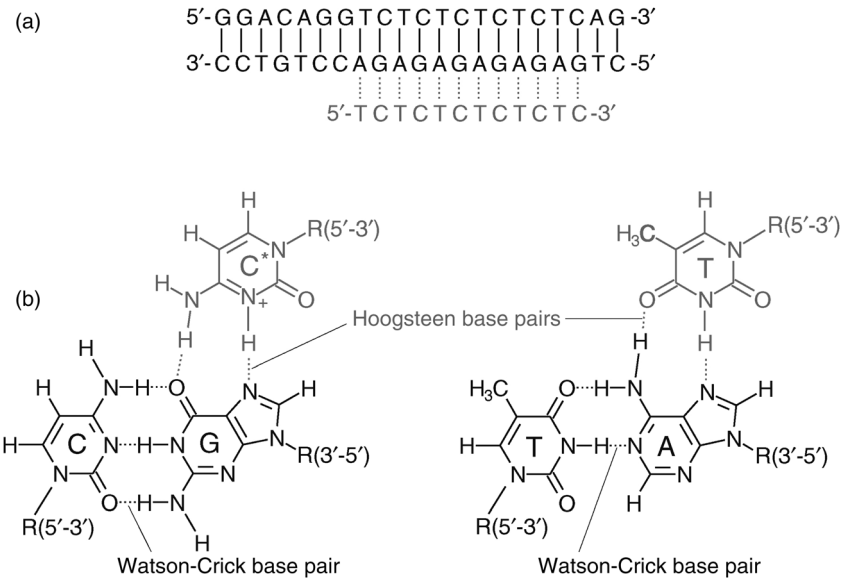


Figure 38.4 (a) Sequence example for an intermolecular triple helix, which is formed by the association of an oligonucleotide with the double-stranded DNA. (b) Two examples of Hoogsteen hydrogen bonding.

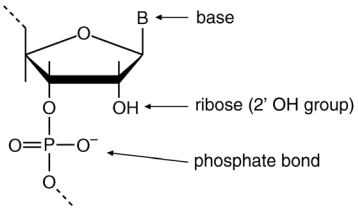


Figure 38.5 Possible positions for modifications of nucleotides.

38.1.3 Modifications of Oligonucleotides to Decrease their Susceptibility to Nucleases

Unmodified RNA- or DNA-oligonucleotides are completely degraded within a few minutes or hours in biological fluids, usually before they can reach the site at which they are intended to work. Therefore, oligonucleotides must be protected from nucleolytic degradation by the incorporation of modified nucleic acids. There are three basic positions where nucleic acids can be changed: the bases, the sugar, or the phosphodiester bond (Figure 38.5). Although it has been shown that modification of the bases serves to increase nuclease resistance, this strategy plays a secondary role in current research. The following section will focus on the more common approaches involving the use of derivatives with modified phosphodiester bonds or a modified sugar.

Numerous nucleases recognize single-stranded oligonucleotides and cleave their phosphodiester bonds. An obvious approach, therefore, is to change these such that they no longer act as a substrate for the nucleases, which stabilizes the oligonucleotides for a long time in serum or in cells. One of the first and most common modifications is the phosphorothioate, in which one of the two oxygen atoms, which are not directly involved in the phosphodiester bond, is replaced by a sulfur atom (Figure 38.6). Oligonucleotides with phosphorothioate bonds are stable for hours in human serum. They hybridize through Watson–Crick base pairing with complementary RNA molecules and induce their degradation by RNase H, like oligonucleotides with normal phosphodiester bonds. However, phosphorothioates have a few disadvantages. For example, their affinity for the target mRNA is less than that of an unmodified DNA oligonucleotide. In addition, oligonucleotides with phosphorothioate bridges bind to a few proteins, particularly those that interact with polyanions such as heparin-binding proteins, and can trigger toxic side effects as a result. The binding of phosphorothioates to albumin in serum of experimental animals or patients also has the advantage that they increase the retention of the oligonucleotides in blood, since they are otherwise rapidly excreted.

Owing to the disadvantages of phosphorothioates, nucleotides are also modified on other positions. A possibility is to link functional groups with the C2′ position of the ribose (Figure 38.6). The most common substituents are methyl or methoxyethyl groups. These modifications also confer a high nuclease resistance on the oligonucleotides. The toxicity of these second-generation oligonucleotides is less than that of phosphorothioates. In addition, their membrane penetration is increased by the linkage with lipophilic substituents and their affinity to the target RNA is increased.

Over the years, nucleic acid chemists have succeeded in improving the properties of oligonucleotides by creating hundreds of DNA analogs with differing properties. A very drastic step is to completely replace the sugar backbone of the DNA. An example is peptide

PNA probes Section 28.2.3

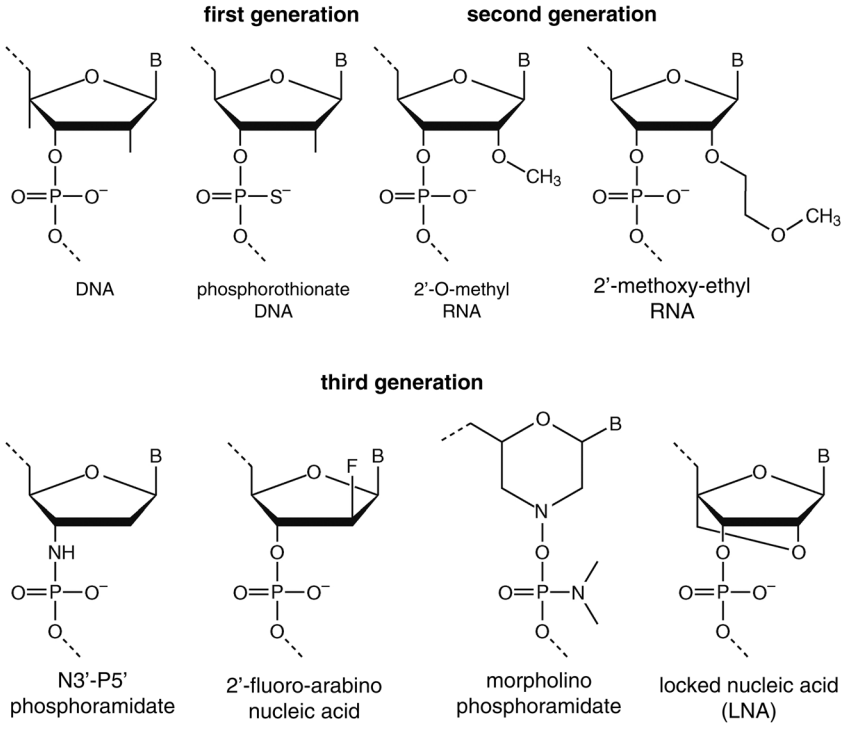


Figure 38.6 Frequently used modified components for antisense oligonucleotides.

nucleic acids (PNAs). In these oligonucleotides, the ribose group and the phosphodiester bonds are replaced by amide bonds, similar to those between amino acids (Figure 38.7). Other examples shown in Figure 38.6 demonstrate the variety in use today. The N3′-P5′ phosphoramidates replace an oxygen bridge with an amino group. In 2′-fluoro-arabino nucleic acid and morpholino-phosphoramidates, the ribose of the nucleotides is replaced by a fluorine-substituted arabinose or even by a six-membered ring. Locked nucleic acids are bicyclic compounds that are conformationally locked by a methylene bridge between the C2′ and C4′ atoms of the ribose. Many of these newer modifications have been used successfully in antisense experiments. They are characterized by high nuclease resistance, high affinity for the target mRNA, and low toxicity.

A serious disadvantage of most oligonucleotides constructed from monomers with modifications to the ribose is that they are not recognized by RNase H as a substrate and therefore

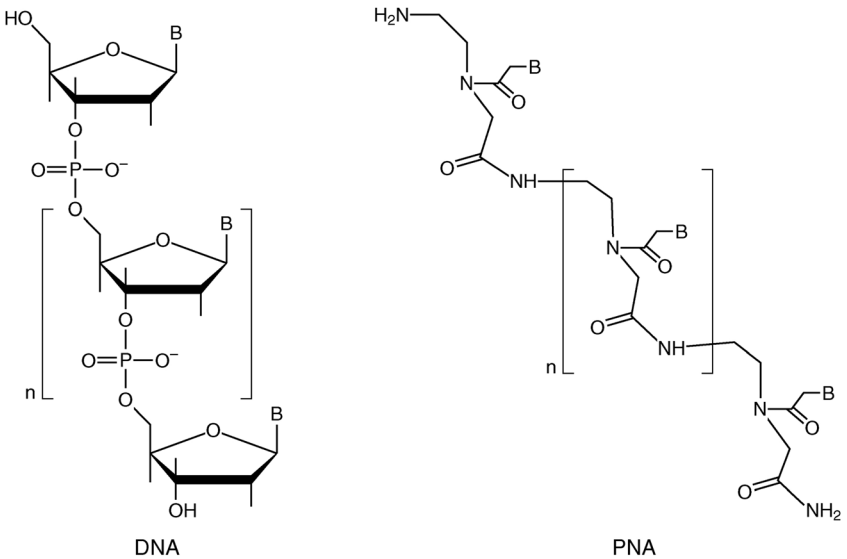


Figure 38.7 Comparison of the backbone of an oligodeoxyribonucleotide (DNA) with that of a peptide nucleic acid (PNA). With this modification, the sugar backbone is replaced by amide bonds.

Figure 38.8 Example of a gapmer. The first five nucleotides on the 5′- and 3′-ends are modified monomers (2′-O methyl RNA or locked nucleic acids), which protect the oligonucleotide against exonucleases. The gray-shaded monomers in the center are unmodified deoxyribonucleotides or phosphorothioates, which guarantee the activation of RNase H.

cannot induce degradation of the target mRNA. A solution is to employ chimeric oligonucleotides that combine different monomers. For example, the ends of the oligonucleotide can be protected from the exonucleases that are dominant in serum by using 2′-O methyl components. A middle piece of 5–8 deoxynucleotides guarantees the induction of RNase H. Such oligonucleotides are referred to as gapmers, due to the "gap" between the protected bases (Figure 38.8). They can be protected from other endonucleases by replacing the phosphodiester bonds by phosphorothioates. Interestingly, the toxicity of the phosphorothioates in this combination is significantly reduced.

38.1.4 Use of Antisense Oligonucleotides in Cell Culture and in Animal Models

Antisense oligonucleotides have been used successfully in many studies in cell culture and animal models. One of the most important hurdles for antisense experiments is the transport of the antisense oligonucleotide to the point at which it works, usually referred to as delivery. The same problem is shared when using other types of oligonucleotides, such as ribozymes or small interfering RNAs, as described below. Oligonucleotides are negatively charged molecules that cross hydrophobic cellular membranes inefficiently. In cell culture experiments cationic lipids are often used as transfection reagents to get the oligonucleotides across the membranes. Remarkably, *in vivo* oligonucleotides are taken up to a certain degree, though the mechanism is unknown. Improving transfection efficiency *in vivo* is essential if oligonucleotides are ever going to become broadly applicable to a variety of diseases. Recent results indicate that antisense oligonucleotides can be drastically shortened, down to 12mers, by using high affinity nucleotides like LNAs. These short oligonucleotides can be used in cell culture without transfection reagents to knock-down the expression of a target gene. This allows them to better mimic the expected results of use *in vivo*, which also do not require the use of additional reagents.

LNA Probes, Section 28.2.4

It is extremely important to prove the specificity of an oligonucleotide in every antisense experiment. The reduction of the expression of the mRNA of the target gene should be shown with Northern blots or quantitative RT PCR and of the protein with Western blots. In addition, negative controls with an independent sequence need to be tested and shown to lack activity. In this way one can be sure that the observed phenotype is caused specifically by the inhibition of gene expression by an antisense oligonucleotide. Unspecific effects can be triggered, for example, by CpG motifs, since they induce an immune response. These effects are undesirable for most antisense applications, but in some special cases are used intentionally, as a component of cancer therapy.

Amplification of RNA (RT PCR), Section 29.2.3

Thousands of studies have described cases in which the expression of target genes has been successfully inhibited with antisense approaches and numerous questions have been investigated this way. While many biochemical techniques are suitable for the investigation of specific questions, such as the analysis of membrane proteins or kinases, a great advantage of the antisense strategies described here is that they are universal, since all mRNAs are structurally similar, even if they code for different proteins. At least theoretically, any desired gene can be downregulated by an antisense approach. For example, antisense strategies are used to investigate the function of closely related kinases that cannot be distinguished with classical biochemical or pharmacological methods. In this way it is possible to determine which isoform is particularly relevant for tumor growth. In other studies, the role of a special receptor in the transmission of signals in different forms of pain was investigated. The special thing about antisense oligonucleotides is that they are not only suitable for basic research, they may also be used as therapeutics. They are suitable for the therapy of any disease that is caused by the expression of a harmful gene.

38.1.5 Antisense Oligonucleotides as Therapeutics

The use of oligonucleotides that are specific only for certain selected mRNAs or DNAs is appealing, since it largely avoids unspecific and unwanted effects often seen for small molecules that can inhibit molecules other than their intended targets and thereby cause

side effects. The modulation of miRNAs also holds promise for modulating entire classes of related mRNAs. For natural selection to work on miRNAs, they must necessarily work on functional groups of mRNAs involved in some phenotype. In cases where the phenotype is disease-relevant, miRNAs represent natural nodes that allow tuning of gene activity.

Human clinical trials with antisense oligonucleotides were begun in the 1990s, after their effectiveness and safety could be shown in experiments in different animal models. The first antisense molecule approved by the US Federal Food and Drug Administration (FDA) in 1998 was used for the treatment of cytomegalovirus-induced retinitis. This progressive, infectious inflammatory disease leads to blindness, most often in immunocompromised AIDS patients. The phosphorothioate oligonucleotide is injected directly in the eye and prevents the loss of vision, conveniently bypassing the usual delivery problem when trying to treat other tissues. It took another 15 years before a second antisense oligonucleotide was approved by the FDA: the 2′-methoxyethyl-modified antisense oligonucleotide mipomersen is directed against apolipoprotein B and is used for the treatment of hypercholesterolemia.

About a dozen antisense oligonucleotides are currently in various stages of clinical development. A few of these molecules are designed to inhibit the replication of viruses. There is a particularly great need for the development of new therapeutic approaches in this area, since many of the currently available antiviral drugs, mostly polymerase inhibitors, are very unspecific and have strong side effects. At the same time, the sequences of many viruses are now known, so that antisense oligonucleotides can be synthesized that are specifically directed against viral RNA.

Antisense oligonucleotides are used in other studies for the treatment of cancer after conventional chemotherapy has failed. Since oncogenes promote the growth of many tumors, inhibiting their expression with antisense strategies is an obvious approach. Other oligonucleotides are used in studies to treat inflammatory diseases, like ulcerative colitis, and heart disease.

The first clinical trials used phosphorothioate antisense oligonucleotides; however, more recent human clinical trials increasingly employ oligonucleotides with modifications of the second and third generations. The expectation that antisense oligonucleotides bind completely specifically to their target sequence in the mRNA without causing unspecific effects on other cellular components has not yet proven to be the case. Nevertheless, the side effects observed in most cases have not been severe. The bigger problem for the use of antisense oligonucleotides is their limited therapeutic efficacy, which is, at least in part, due to delivery issues.

Cytomegalovirus DNA virus of the family Herpesviridae, which is endemic in most human populations. It may lead to blindness in immunocompromised patients.

38.2 Ribozymes

38.2.1 Discovery and Classification of Ribozymes

Until the beginning of the 1980s, it was assumed that only proteins possess enzymatic activity. Then Thomas Cech and coworkers discovered that certain RNA molecules of the protozoan *Tetrahymena thermophila* splice themselves. This means that even in the absence of proteins these RNAs have catalytic ribonuclease activity. Cech named ribonucleic acids with enzymatic activity ribozymes. A short time later, the group led by Sidney Altman discovered that the ubiquitous RNase P, which processes the 5′ ends of tRNAs, is a ribozyme. RNase P is a complex of RNA and protein; interestingly, the RNA is catalytically active, even without the protein component. An important feature of this ribozyme is the fact that it does not process itself, but instead processes an independent substrate. It emerges unchanged from the reaction and thereby fulfills the classic definition of an enzyme. Cech and Altman were awarded the Nobel Prize in Chemistry in 1989 for their revolutionary discovery.

In the meantime, different types of ribozymes are known that can be roughly categorized into those of a few hundred to 3000 nucleotides and small ribozymes of 30–150 nucleotides in length. Particularly well studied are the hammerhead ribozymes, which were discovered in plant pathogens as self-splicing RNAs. For practical use it was important to develop a variant that is capable of catalyzing an independent substrate. The hammerhead ribozyme shown in Figure 38.9a consists of a catalytic center of 22 nucleotides and two substrate recognition arms of seven to nine nucleotides, which bind the target RNA. The binding arms can be so designed that they can recognize any chosen sequence. In a metal ion-dependent step, the target RNA is

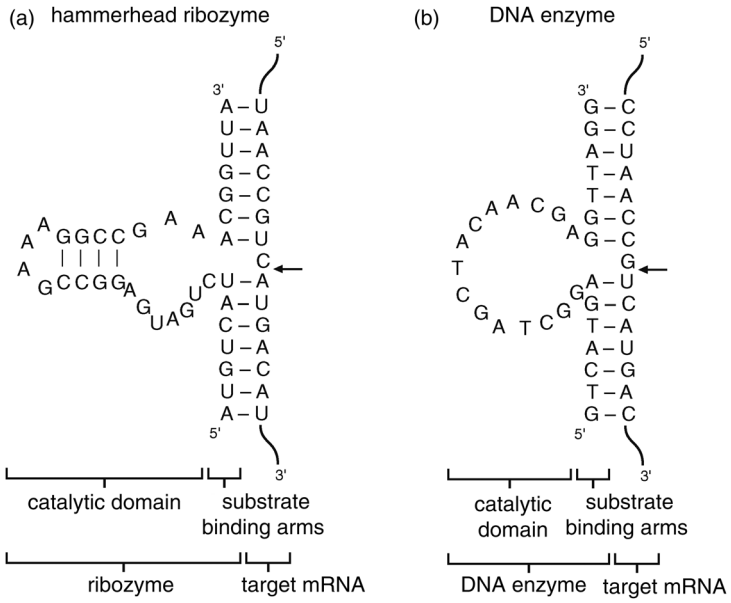


Figure 38.9 Structure of a hammerhead ribozyme (a) and an RNA-cleaving DNA enzyme (b). The arrows show the position of the cleavage sites.

cleaved and the two cleavage products diffuse away, so that the ribozyme is available for cleavage of further RNAs (Figure 38.10).

In recent years catalytic activities of ribozymes in other biochemical processes have been discovered. The elucidation of the structure of the ribosome showed that there are no proteins near the catalytic center. This includes a central process of cells, the peptidyl transferase reaction during protein synthesis, which is catalyzed by a ribozyme. However, it is not definitively established whether protein components might still be involved in some way in the coupling of amino acids. Ribozymes have also been found in the 5′ untranslated region of mRNAs in bacteria. These are regulated by metabolites and thus control the expression of genes that are important for metabolism. Finally, a ribozyme has been discovered that catalyzes the separation of RNA polymerase II from the mRNA during termination of the transcription of the globin gene in eukaryotes. The discovery of more and more new ribozymes supports the RNA world hypothesis. This hypothesis holds that before the current world dominated by DNA and proteins, life forms existed in which RNA played the most important role, since RNA can both transmit inheritable information and catalyze reactions.

38.2.2 Use of Ribozymes

Similar to the selection of antisense oligonucleotides, there are a few important issues in the selection of ribozymes. Different ribozymes have differing preferences for cleavage sites;

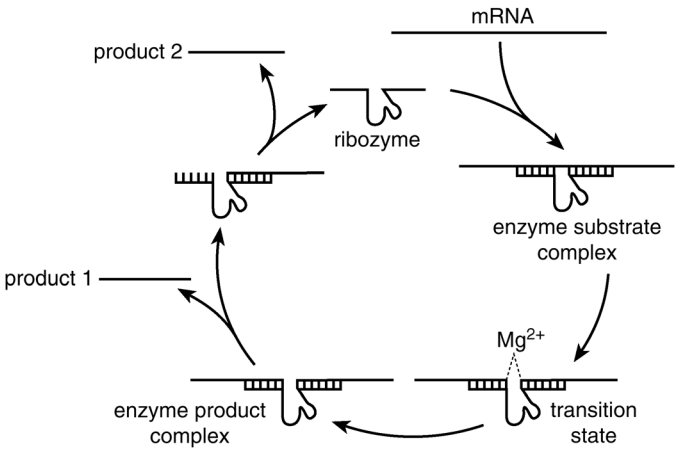


Figure 38.10 Catalytic cycle of ribozymes: The ribozyme binds by sequence-specific Watson–Crick base pairing to its complementary target mRNA. In the presence of metal ions, the reaction moves through a transition state, resulting in the cleavage of the substrate. The two product strands are set free and the ribozyme is available for another round of target RNA cleavage.

hammerhead ribozymes cleave after GUC or AUC triplets particularly efficiently. These cleavage sites must, in addition, be readily accessible to the ribozymes, just like the binding region of antisense oligonucleotides. Since ribozymes consist of RNA components, they are more sensitive to nucleases than DNA oligonucleotides. In systematic studies, ribozymes were protected from degradation by the incorporation of modified nucleotides, primarily 2′-O-methyl RNA. To preserve the catalytic activity, unmodified ribonucleotides must remain in five positions. This lead to the idea of creating enzymes from DNA, which is less susceptible to nucleolytic degradation in biological media than RNA. This succeeded with the help of a combinatorial approach, as described in Section 38.4 about aptamers. The resulting DNA enzyme (Figure 38.9b) cleaves target mRNAs very efficiently. It is inexpensive and easily synthesized and is more resistant to nucleases than unmodified hammerhead ribozymes.

Not only can ribozymes composed of RNA be chemically synthesized and transfected into cells or injected into animals (exogenous application), they can also be transcribed directly in cells (endogenous expression). The DNA sequence that codes for the ribozyme is cloned into a vector, which is introduced into the target cell. The ribozymes are then synthesized continually within the cell.

Ribozymes have been successfully employed in animal models since the 1990s. Similar to antisense oligonucleotides, they have been tested in various indications, covering viral infections, cardiovascular disease, cancer, and rheumatoid arthritis. Some ribozymes have been tested in the early phases of clinical trials on patients. Patients infected with HIV had blood withdrawn and transfected *ex vivo* with retroviral ribozyme vectors and subsequently readministered to the patients. The intent of this experiment was to make the lymphocytes resistant to HIV. In other studies, chemically synthesized ribozymes with modified nucleotides were given to hepatitis C or cancer patients. The ribozymes were usually well tolerated in the clinical trials but their effectiveness was inadequate. A possible explanation is the dependence of the ribozyme catalyzed reactions on metal ions. In most *in vitro* studies the concentration of magnesium ions is significantly higher than the intracellular concentration. Presumably the versions of the hammerhead motifs used in most studies as shown in Figure 38.9 lacked important peripheral regions. These regions are necessary for a high catalytic activity of the natural ribozyme at low magnesium ion concentrations.

38.3 RNA Interference and MicroRNAs

38.3.1 Basics of RNA Interference

A particularly promising method for the specific inhibition of the expression of a gene is RNA interference (RNAi). In 1998, Andrew Fire and Craig Mello discovered that double-stranded RNA molecules in the nematode *Caenorhabditis elegans* can be used to sequence-specifically silence genes. In the meantime the mechanism has been elucidated in detail (Figure 38.11): the double-stranded RNA is first cleaved by an RNase, Dicer, into short RNA duplexes, which are called small interfering RNAs (siRNA). Next the siRNA is incorporated into a protein complex, the RNA-induced silencing complex (RISC), during which one of the two strands is discarded. The other strand guides RISC to the target mRNA and hybridizes with it by conventional base pairing. RISC contains an endonuclease that cleaves the mRNA at a defined position. The protein complex containing the antisense strand of the siRNA then dissociates from the substrate and becomes available to initiate a new cleavage. The cleaved ends of the mRNA are no longer protected by a cap or poly A tail and are quickly degraded by cellular RNases. Similarly to the use of antisense oligonucleotides and ribozymes, the synthesis of the protein is inhibited.

Interferon response Interferons are species-specific proteins with antiviral and immunomodulatory properties.

The phenomenon of RNAi is an evolutionarily conserved mechanism whose natural function is still not completely understood. There is evidence that they serve to defend against viruses and to protect cells from mobile genetic elements, such as transposons. After their discovery, RNAi was used intensively for research in model organisms such as *Caenorhabditis elegans* and *Drosophila melanogaster*. The method could not be used in mammalian cells, however, because double-stranded RNA triggers an interferon response, which leads to a general block of protein expression. It was therefore an important

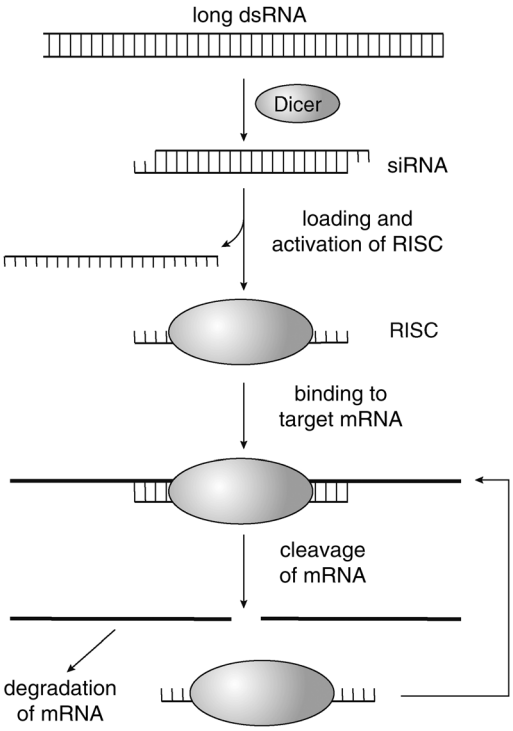


Figure 38.11 Mechanism of RNA interference.

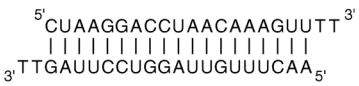


Figure 38.12 Typical example of a small interfering RNA (siRNA).

breakthrough when Thomas Tuschl and his coworkers showed that short, 21–23 nucleotide long siRNA molecules sequence-specifically block gene expression in mammalian cells, since the interferon response is only triggered by double-stranded RNA molecules that are over 30 nucleotides long. Figure 38.12 shows a typical example of a siRNA: it consists of two 21 nucleotide long strands, which form a 19-mer duplex. Each end usually has two overhanging nucleotides, which in most cases are deoxythymidine.

For the practical use of RNAi, an efficient siRNA must first be generated, which requires a sophisticated design based on particular sequence criteria, such as the GC content and the relative stability of the two ends of the complex. In addition, the structure of the target mRNA plays a role. Most researchers assume that a good siRNA is significantly more potent than an antisense oligonucleotide or a ribozyme, which means that a lower concentration is required to turn the target gene off.

Although RNA interference can be seen as a sequence-specific and very efficient method to post-transcriptionally inhibit gene expression, it is important to remember that it can also trigger unspecific side effects. Under certain conditions siRNAs can inhibit partially homologous RNAs other than their intended targets. In addition, in some cases siRNAs, depending on their sequence, trigger an unspecific interferon response and also induce other pathways of the innate immune system, among others those downstream of Toll-like receptor 3. At high concentrations, the RNAi-triggered cellular processes can disturb the microRNA pathway described below, which can have toxic effects. These results show that RNAi experiments need to be carefully planned, including numerous controls, the results need to be interpreted with great care, and particular caution must be exercised for therapeutic approaches.

38.3.2 RNA Interference Mediated by Expression Vectors

The double-stranded siRNAs are unexpectedly stable for RNA molecules, since they are presumably protected from nucleolytic degradation by proteins. They can be stabilized further by the incorporation of modified nucleotides (Section 38.1.3). However, even with these measures, the effects of chemically synthesized siRNAs are transient and only last a few days. A longer lasting blockade of gene expression can be accomplished with special plasmids encoding expression cassettes for short hairpin RNAs (shRNAs) (Figure 38.13). These self-complementary RNAs are expressed intracellularly and subsequently processed to a typical siRNA by the enzyme Dicer, as described above. The shRNA is coded by the expression vector and

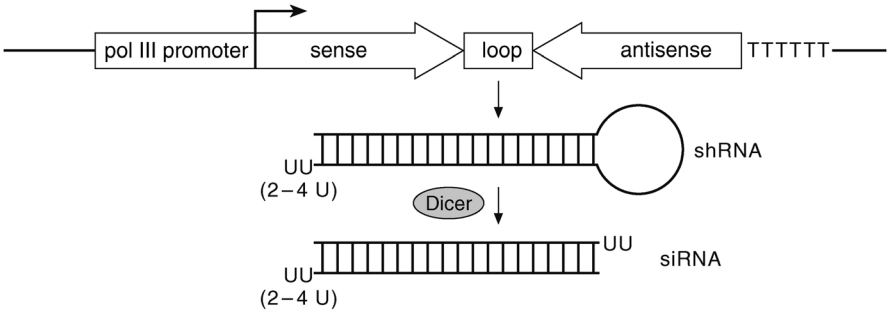


Figure 38.13 Vector expression of short hairpin RNA (shRNA). The shRNA is transcribed under the control of polymerase III promoters and intracellularly processed to siRNA.

transcribed under the control of an RNA polymerase III promoter, which is suitable for the expression of short RNAs without a cap structure or poly A tail.

Owing to the continuous intracellular transcription of the double-stranded RNA, its effects are relatively long lasting. It is even possible to stably transfect eukaryotic cells with such an expression vector so that the siRNA is always present. In addition, the expression cassettes, consisting of a promoter and the shRNA coding sequence, can be incorporated into viral vectors. Such constructs have already been used in clinical trials. In particular, oncoretro- and lentiviruses, adenoviruses and adeno-associated viruses are used, which are modified to efficiently transduce a therapeutically effective gene into cells and, for safety reasons, are incapable of replicating themselves or spreading. Viral vectors can be used for the transduction of cells for research purposes; in addition, they offer the promise of the therapeutic application of RNA interference.

38.3.3 Uses of RNA Interference

RNA interference (RNAi) is one of the most important developments in molecular biology since the development of PCR. In only a few years it has advanced to become a widespread standard method that has been employed in thousands of publications. The uses of RNAi range from functional studies of individual genes to genome wide screens to therapeutic applications. For many research projects, the function of individual genes is studied by inhibiting its expression. In this case RNAi is significantly faster than antisense oligonucleotides or ribozymes. For example, genes that are suspected of playing a role in cancer can be turned off by siRNA. The effect on cell growth can subsequently be studied in cell culture or animal models.

A completely new possibility opened by RNAi is carrying out genome-wide screens, in which every gene in an organism is turned off, one at a time. In this way, all the genes can be identified that are involved in a cellular or pathological process. Libraries have been created for both the human genome and model organisms that consist of siRNAs or vector-encoded shRNAs against every individual gene. Such screens have been used, for example, to identify hundreds of host factors that HIV-1 requires for infection and replication. In other experiments, previously unknown proteins have been identified that are responsible for the uncontrolled proliferation of tumor cells. After elucidation of the human genome, RNAi has provided a method that allows determination of the function of the encoded proteins.

A further step to the characterization of a gene is its investigation in animal models. However, for this purpose the biggest problem in the use of RNAi must be solved, namely, the efficient delivery of the siRNA into the cells of the target tissue. As a result, many methods have been developed for the transfer of oligonucleotides. For example, siRNAs can be brought into the cells with positively charged lipids or nanoparticles, similar to what has been described for antisense oligonucleotides. Alternatively, lipophilic molecules like cholesterol can be coupled directly to the siRNA to ease passage across the membrane. If an siRNA needs to be specifically targeted to just one cell type, antibodies and aptamers (see Section 38.4 below), which recognize specific markers on the cells, can guide them to cells and facilitate entry. In addition, viral vectors can be used for RNAi approaches. As previously described, shRNA expression cassettes can be transfected efficiently into cells with the help of these vectors. Through the selection of suitable capsid proteins, it is possible to optimize this process for special target tissues.

RNAi developed into a standard method for the analysis of gene function in molecular biology laboratories very rapidly. At the same time its potential as a technology for therapeutic use was recognized. Only a few years after the discovery of RNAi, clinical trials were begun, some of which

have already reached advanced phases. The first treatments have been for diseases that can be treated with local application, such as the eye, to get around the delivery problem. However, most of these trials were terminated due to unspecific modes of action of the siRNAs. In the meantime clinical tests have also been initiated in other fields such as cancer, virus infections, and metabolic disease. These approaches involve systemic delivery and the application of viral vectors.

For example, a clinical trial investigating the potential of RNAi as a new therapeutic strategy aims at treating infections with the respiratory syncytial virus (RSV). While this virus produces mild symptoms in healthy adults, it can lead to severe, and in many cases fatal, complications in newborns and premature babies. RSV infects the respiratory tract, so that the therapeutic siRNA can simply be delivered by inhalation. Tests with infected adults resulted in significant antiviral effects.

More than two dozen clinical trials based on RNAi have been initiated in the meantime. Most of them aim at treating eye disease, virus infections, and cancer. In general, the RNAi treatment was found to be well tolerated. Now the therapeutic benefits have to be demonstrated in later phases of clinical development.

38.3.4 microRNAs

The natural match to siRNAs are intracellularly expressed, 21–23 nucleotide long RNA molecules, which are called microRNAs (miRNAs). The human genome codes for almost 2000 miRNAs. The miRNAs are transcribed as part of a longer pri-miRNA (Figure 38.14). While still in the nucleus, the long transcript is processed by the RNase Drosha to the roughly 70 nucleotide long pre-miRNA, which is exported into the cytoplasm by exportin-5. Once there, Dicer performs the final cleavage to the mature form. The fully processed miRNA is loaded into RISC, like the siRNAs. The exact mechanism of the miRNA-induced post-translational repression of gene expression has not been completely elucidated and they possibly work by more than a single mechanism. When highly complementary to the target RNA they induce its cleavage and degradation, like siRNAs. They can work, however, even when they are only partially complementary to their target RNA. In this case, they repress translation without causing the cleavage of the mRNA. In addition, miRNAs can destabilize the target mRNA by

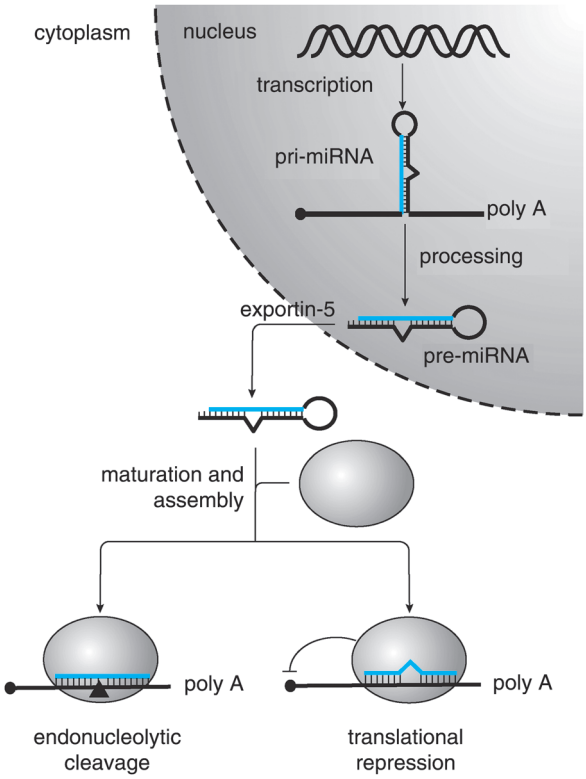


Figure 38.14 Schematic of the micro-RNA pathway described in the main text.

inducing the removal of the protective poly A tail and the cap at the 5′ end. The preferred binding region for miRNAs is the 3′ untranslated region of target mRNA.

According to current results, miRNAs control the expression of more than 60% of all protein-coding genes. It is therefore not surprising that they are involved in almost every cellular and pathological process investigated so far. In a first example of the importance of miRNAs in physiological processes, it was shown that they control the expression of genes that trigger the release of insulin. But miRNAs are also involved in numerous disease processes. In many types of tumor miRNAs are deregulated, which can lead to a dysregulation of genes involved in apoptosis or control of the cell cycle. There is also a complex relationship between cellular miRNAs and viruses, which will be described in more detail below.

To investigate the relevance of miRNAs in a disease process, the miRNA levels of cells in diseased tissue is compared with the levels in healthy control cells. DNA arrays containing a complete collection of probes complementary to all miRNAs can be used for this purpose. Alternatively, the deep sequencing technique can be used. By massive sequencing of the short RNAs conclusions can be drawn about the frequency of a miRNA and the level of its expression. With this method, some miRNAs are usually identified that are particularly strongly or weakly expressed in a disease process. These results can be verified with different methods. A frequently used technique for this purpose is quantitative RT PCR. But, in contrast to conventional quantitative RT PCR, only very short, 21–23 nucleotide long RNAs need to be reverse transcribed and amplified. Therefore, reverse transcription usually employs a stem loop primer that lengthens the fragment. Systems have been developed that allow analysis of the complete pattern of expression of all known human miRNAs by this technique. Other methods that can be used to investigate the expression level of miRNAs are Northern blots, which must also be adapted to hybridization with very short RNAs, and RNase protection assays. Surprisingly, miRNAs can even be detected in the circulation. They are therefore seen as potential biomarkers for various cellular states and diseases. For example, it has been shown that the level of miRNA-208a, detected with quantitative RT PCR, increases after a heart attack. A corresponding test would allow the early diagnosis of myocardial infarction.

Quantitative PCR, Section 29.2.5

Many miRNAs are expressed in a tissue-specific manner. For example, miRNA-122 is exclusively expressed in the liver. Owing to their high affinity for complementary RNA, LNA modified probes are particularly suitable for the detection of this sort of tissue specificity by *in situ* hybridization. Such a detection can be carried out in a whole organism, for example, transparent zebrafish embryos.

The identification of the cellular targets of the miRNAs remains a major challenge. Various internet tools exist to determine the potential target RNAs for a given miRNA. However, exclusively bioinformatic approaches still have a high failure rate. Genome or proteome wide investigations of the regulator effects of a miRNA with DNA arrays and proteomic methods are still, in comparison, very labor-intensive.

In general the function of upregulated miRNAs is investigated in a further step with the aid of complementary antisense oligonucleotides. Modified antisense oligonucleotides, which bind with high affinity to the miRNA, are particularly useful for this purpose. This approach reduces or eliminates the effects of the miRNA, which allows conclusions to be drawn about their role in cellular processes. For example, it has already been shown that replication of hepatitis C virus is dependent on the expression of the previously mentioned liver-specific miRNA-122. If the miRNA is blocked by an antisense oligonucleotide, the virus can no longer proliferate. This result was used in the first clinical trial involving miRNAs, in which an LNA-modified antisense oligonucleotide against miRNA-122 was used to treat hepatitis C infections. If a miRNA is underexpressed in a pathophysiological process, chemically synthesized miRNA or endogenous expression of same can be used to balance the level.

38.4 Aptamers: High-Affinity RNA- and DNA-Oligonucleotides

38.4.1 Selection of Aptamers

In the techniques described up to this point, the critical property of oligonucleotides was their ability to hybridize with complementary sequences via Watson–Crick base pairing.

Aptamers and the SELEX Procedure, Section 32.9.2

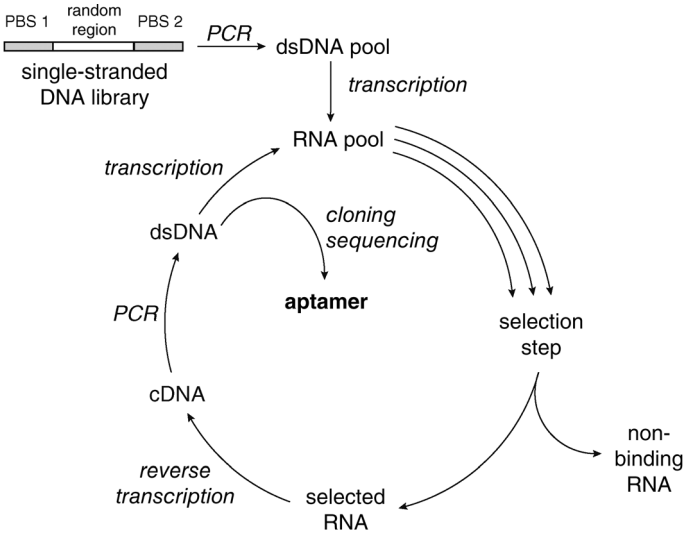


Figure 38.15 SELEX-strategy for the isolation of RNA aptamers with high affinity. PBS: primer binding site.

Oligonucleotides form, however, complex three-dimensional structures through which they can bind to target molecules with high affinity, similar to antibodies. In the early 1990s a procedure was developed that allowed selection of those structures of a large pool of oligonucleotides that specifically bind to a given target. These high-affinity oligonucleotides are called aptamers and can be directed against all sorts of target structures, such as ions, small organic molecules, amino acids, nucleotides, proteins, or even whole viruses or cells.

The combinatorial approach used to isolate aptamers is called systematic evolution of ligands by exponential enrichment (SELEX). It consists of a cyclical process in which the sought-after molecule is successively enriched in the population (Figure 38.15). To begin with a combinatorial library of single-stranded DNA molecules is chemically synthesized. These oligodeoxyribonucleotides consist of two flanking, fixed sequences that serve as binding sites for PCR primers and a random section between them, which is the basis for the variety of the oligonucleotide mix. The theoretical complexity of a library of oligonucleotides with n randomized positions is 4^n (e.g., for a random sequence section that is 25 base pairs long $4^{25} \approx 10^{15}$ different sequences are possible that can form different structures).

PCR, Chapter 29

First the oligodeoxyribonucleotides are amplified by PCR to generate double-stranded DNA molecules and to replicate the library. Owing to the additional hydroxyl group, RNA oligonucleotides were initially thought to be able to form more complex three-dimensional structures, so that DNA libraries are usually transcribed with T7 RNA polymerase into a mix of RNA oligonucleotides. However, more recently it was shown that DNA aptamers with similar properties and target affinities as RNA aptamers can be selected. Simple strategies solely based on DNA molecules have become the most common approach. The decisive step is the selection of those oligonucleotides that, due to their structure, bind particularly tightly to the target. There are different procedures for this; for example, a protein can be bound to a solid support in a column and the oligonucleotide mix is run through the column. Non-binding oligonucleotides pass through the column, while those with a high affinity to the target remain in the column. They are then eluted after repeated washing. They are now enriched; however, only a very small amount is available so that additional amplification steps are necessary. This is accomplished by reverse transcribing the RNA oligonucleotides into DNA oligonucleotides and amplified with PCR, followed by further rounds of selection. After several SELEX rounds have been completed the strongly binding oligonucleotides are highly enriched in the mixture, so that they can be cloned and sequenced. Sequence families are identified through bioinformatic analysis and the structures are predicted. Finally, the oligonucleotide can be optimized through further evolutionary procedures (e.g., reselection with intentionally error-prone PCR) and by shortening. The result of this process is an aptamer with a high affinity to the chosen target molecule. Besides this *in vitro* procedure, other methods have been developed to select aptamers against cell surface components of whole cells in culture or even tumor markers in living animals.

Aptamers are comparable to antibodies in terms of their function, binding to a target molecule. Both types of molecule have similar binding constants; however, aptamers possess a

few advantages: antibodies are typically generated by immunizing animals. This can cause problems if the protein is not immunogenic and fails to trigger antibodies. The standard SELEX procedure is, in contrast, a purely *in vitro* procedure in which aptamers can be generated against any chosen target molecule, at least in principle. The aptamers can be reproduced at any time, easily and in high purity, by chemical synthesis. They are larger than small molecule compounds, but usually about an order of magnitude smaller than antibodies, so that they penetrate dense tissue, such as in a tumor, better than antibodies. Another advantage is their high specificity. By a targeted procedure, a so-called counter-SELEX, aptamers are obtained that can differentiate between very closely related chemical structures (e.g., theophylline and caffeine). In addition, SELEX has been standardized to such a degree that it can be carried out automatically by pipetting robots. This allows aptamers to be generated quickly and in parallel. For membrane proteins, aptamers selected by *in vitro* SELEX sometimes do not recognize their target in the cellular context, since the protein folds differently in its natural environment. To overcome these problems, sophisticated strategies to select aptamers against membrane proteins in whole cells have been developed. In addition, even the selection of aptamers against tumors in living animals has been carried out successfully.

So far, *in vivo* testing of aptamers has not found them to be toxic or immunogenic. However, as for other oligonucleotides, aptamers must also be protected against nucleases. Regions of the aptamer, which are not directly involved in the binding of the target molecules, can be stabilized with modified nucleotides, but the regions directly involved in binding cannot be modified this way after SELEX without risking loss of affinity. As a result, the modifications are often employed during the SELEX. They must, however, be accepted by the enzymes employed in SELEX (RNA polymerase, reverse transcriptase, DNA polymerase), which severely limits the selection. Frequently, monomers carry an amino group or a fluorine atom on the 2′ position of the ribose (Figure 38.16). An alternative is the use of Spiegelmers, which consist of L-RNA instead of the naturally occurring D-RNA. These enantiomeric monomers are not recognized by nucleases, so that they are extremely stable. However, their selection requires a sophisticated selection procedure in which first a normal aptamer is generated against a mirror image of the target molecule. The enantiomer of this aptamer is then the sought-for Spiegelmer, which then binds to the normal target.

As described above, a major challenge for the application of oligonucleotide-based strategies is their intracellular delivery. Aptamers differ from antisense oligonucleotides, ribozymes, and siRNAs, in that they can also be directed against extracellular targets such as growth factors and other signaling molecules in the blood stream thereby avoiding the need to cross the cell membrane. However, aptamers can also be used to investigate the intracellular function of proteins. In this case, the aptamer either has to be transferred into the cell or it can be transcribed intracellularly by an expression cassette.

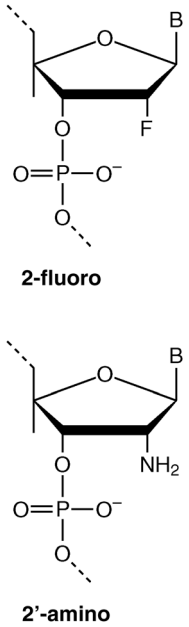


Figure 38.16 Modifications of the 2′ position of the ribose by fluorine (a) or an amino group (b), which are used to stabilize aptamers.

38.4.2 Uses of Aptamers

Owing to the large number of possible target molecules for aptamers, they can be used in many applications. Even the potential of aptamers as therapeutics has been investigated in clinical trials. A particularly successful example was one that treats age-related macular degeneration. It is directed against a growth factor that controls the growth of blood vessels, the cause of the disease. This aptamer was approved at the end of 2004 by the FDA as a drug for the treatment of age-related macular degeneration. Aptamers for the treatment of cancer and for the inhibition of blood coagulation have also been tested clinically. Roughly a dozen aptamers were or are being tested in clinical trials. High-affinity oligonucleotides can be used for diagnostic purposes, as well as therapeutic. Aptamers that recognize tumor markers expressed in the extracellular matrix have been developed. If a radioactive isotope is coupled to such an aptamer, the precise position of the tumor can be determined.

Besides applications directly applicable to medicine, aptamers have found many other biotechnological uses, which can only be mentioned briefly here. One application involves coupling high-affinity oligonucleotides to the surface of chips and sensors. These detection systems are used for various purposes, from protein analysis to the testing of soil samples for the presence of toxins. Interestingly, aptamers can also be used to control ribozymes allosterically. The aptamer is coupled to a hammerhead ribozyme with a spacer. Binding of the ligand to the

aptamer triggers a conformation change that activates or inactivates the ribozyme. These allosterically regulated ribozymes, called aptazymes, have proven to be very sensitive sensors.

Aptamers are not only artificial products of biotechnology, they are also naturally occurring regulatory systems. These biological sensors are called riboswitches and are usually found in the 5′ untranslated region of an mRNA of various genes. They "measure" the level of a metabolite and regulate the expression of genes through conformational changes. If the product of an encoded gene is present at high concentrations, protein biosynthesis is inhibited. With the aid of analogs of the ligand, the function of the riboswitch can be analyzed more precisely and, possibly, the growth of pathogenic bacteria can be inhibited in this manner.

38.5 Genome Editing with CRISPR/Cas9

Another oligonucleotide-based method with great potential in molecular biology is the CRISPR/Cas9 technology. In prokaryotes, Clustered Regulatory Interspaced Short Palindromic Repeats (CRISPR) and their associated *cas* genes serve as an adaptive immune system to protect them from infection by bacteriophages. The bacterial system was adapted for use in eukaryotic cells and can be used for precise genome engineering.

Cas9 is a DNA endonuclease with two active sites that each cleave one strand of double-stranded DNA. For technological applications, a guide RNA (gRNA) is designed that is complementary to the target region (Figure 38.17). However, the selection of the target is not fully free, as CRISPR/Cas9 systems require a short sequence, called the Protospacer Adjacent Motif (PAM), which is located directly adjacent to the target sequence. For the most widely used Cas9 protein from *Streptococcus pyogenes* the PAM sequence is NGG. Another element of the gRNA is the crRNA (for CRISPR RNA) which interacts with the Cas9 protein (the hairpin in Figure 38.17). Similar to the double-stranded RNA triggering RNAi, the gRNA can either be synthesized chemically or expressed intracellularly.

Oligonucleotide Synthesis, Section 27.6.1

Many applications of the CRISPR/Cas9 technology have been developed in recent years. In the simplest version, the gRNA directs the Cas9 protein to a chosen target sequence and induces its cleavage. The double-strand break will be repaired by a cellular mechanism known as Non-Homologous End Joining (NHEJ). This process, however, is error-prone and often leads to the introduction of mutations which inactivate the targeted gene. This approach can be used to produce and study loss-of-function phenotypes or to inactivate deleterious genes in therapeutic applications. A more sophisticated strategy aims at introducing point mutations or even larger fragments of genetic material. In this case, a DNA fragment with homologous ends to the target region is required, in addition to the gRNA and the Cas9 protein. Following the double-strand break introduced by the Cas9 protein, the homologous DNA fragment is inserted by the Homology-Directed Repair (HDR) pathway. For further applications of the CRISPR/Cas9 technology that repress or activate transcription or influence other cellular processes, the interested reader is referred to the references given at the end of the chapter.

Compared to the antisense and RNAi approaches discussed above, the CRISPR/Cas9 technology is extremely efficient and easy to use. In addition, the CRISPR/Cas9 system enables full knockout of the target gene while the other methods only result in partial

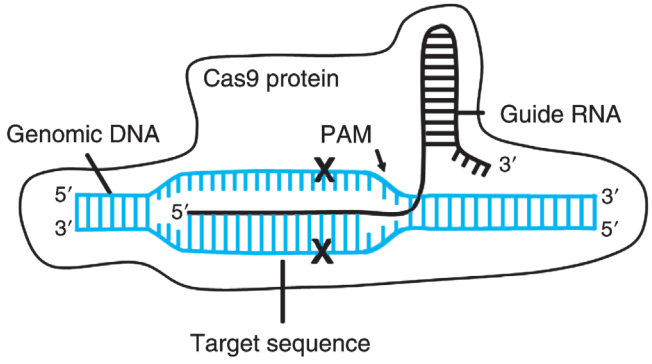


Figure 38.17 Genome editing with the CRISPR/Cas9 system. A guide RNA directs the Cas9 protein to the target sequence in the double-stranded DNA. The endonuclease then cleaves both strands of the genomic DNA. X indicates cleavage site.

knockdown of gene expression. Finally, some studies suggest that the CRISPR/Cas9 system produces less off-target effects than the other methods.

Within only a few years, the CRISPR/Cas9 technology has already found its way into thousands of research laboratories worldwide and has become an indispensable research tool. In addition, several clinical trials involving CRISPR/Cas9 have already been initiated and several more have been announced for the near future. As the technology makes gene editing in germlines technologically viable, its employment in this manner raises serious bioethical concerns.

38.6 Outlook

Antisense approaches possess a strong appeal due to a simple idea that is inherent in their nature: theoretically any given target RNA in a cell can be specifically and selectively blocked or destroyed by the use of complementary oligonucleotides. These techniques have been used successfully for many years in experiments in which the function of a gene is determined by studying what happens when it is inhibited. Antisense strategies often achieve desired results faster than the development of small molecule inhibitors of proteins or the generation of knock-out animals; they are used by pharmaceutical companies for target validation, before the real drug is sought in high-throughput screening of a large compound collection. Since the antisense principle is universally applicable, it can also be used against genes whose corresponding proteins cannot be addressed by small molecules due to their size, structure, or tissue distribution, often referred to as non-druggable targets.

In many cases, however, the practical application of antisense oligonucleotides and ribozymes encounters significant difficulties and the results of clinical studies have often been disappointing. Apparently there are problems in getting enough molecules into the cells of the correct tissue at the right time. Great hopes are placed, as a result, in the newer method of RNA interference, which appears to be significantly more efficient. It has already been employed with great success in numerous studies and valuable new discoveries have been made in large screening campaigns in the areas of virology and tumor biology. Whether small interfering RNAs also prove suitable for therapeutic use remains to be seen. Then the vision could become reality, that the development of a therapeutic can be significantly accelerated by use of RNAi. New potential targets could be identified for the treatment of a disease by employing large libraries of siRNAs. The siRNA used to identify the target can be used for target validation and further developed as a therapeutic without having to begin the search for a low-molecular drug anew.

The significance of miRNAs in the fine regulation of cellular processes is becoming increasingly evident. It is astounding that such a central regulatory mechanism of the cell could remain undiscovered for so long. Many of the most modern technologies of bioanalysis play a role in the investigation of the function of miRNAs. Deep sequencing, array technologies, and quantitative PCR are used to determine the expression level of the miRNAs. Bioinformatics and modern proteomic methods are employed to determine the target mRNAs of the miRNAs and oligonucleotides can be used to inhibit the miRNAs, in order to investigate their function or to correct their deregulation in disease processes.

The main challenge to all antisense-based strategies, the intracellular delivery of the oligonucleotides, can be avoided by using aptamers, since these can be directed against extracellular targets. Aptamers can also be used *ex vivo* for various purposes, from protein analysis to the monitoring of samples for toxins or biological warfare agents. In the coming years we can expect to see an expansion of the areas in which oligonucleotides for diagnostic and therapeutic purposes are used.

Recently, the CRISPR/Cas9 technology has been developed as a new approach for precise RNA-directed genome editing. It has great potential in biomedical research as it is an easy method to disrupt or modify targeted genes. In addition, it will soon become a new therapeutic option for genetic disorders and viral infections.

Further Reading

Bennett, C.F. and Swayze, E.E. (2010) RNA targeting therapeutics: molecular mechanisms of antisense oligonucleotides as a therapeutic platform. *Annu. Rev. Pharmacol. Toxicol.*, **50**, 259–293.

Burnett, J.C. and Rossi, J.J. (2012) RNA-based therapeutics: current progress and future prospects. *Chem. Biol.*, **19**, 60–71

Doudna, J.A. and Lorsch, J.R. (2005) Ribozyme catalysis: not different, just worse. *Nat. Struct. Mol. Biol.*, **12**, 395–402.

Fabian, M.R. and Sonenberg, N. (2012) The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nat. Struct. Mol. Biol.*, **19**, 586–593.

Haussecker, D. (2012) The business of RNAi therapeutics in 2012. *Mol. Ther. Nucleic Acids*, **1**, e8.

Hille, F. and Charpentier, E. (2016) CRISPR-Cas: biology, mechanisms and relevance. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **371**, pii: 20150496.

Jackson, A.L. (2012) Developing microRNA therapeutics: approaching the unique complexities. *Nucleic Acid Ther.*, **22**, 213–225.

Keefe, A.D., Pai, S., and Ellington, A. (2010) Aptamers as therapeutics. *Nat. Rev. Drug Discovery*, **9**, 537–550.

Kurreck, J. (ed.) (2008) *Therapeutic Oligonucleotides*, Royal Society of Chemistry Publishing, Cambridge.

Kurreck, J. (2009) RNA interference: from basic research to therapeutic applications. *Angew. Chem., Int. Ed.* **48**, 1378–1398.

Lightfoot, H.L. and Hall, J. (2012) Target mRNA inhibition by oligonucleotide drugs in man. *Nucleic Acids Res.*, **40**, 10585–10595.

Mulhbacher, J., St-Pierre, P., and Lafontaine, D.A. (2010) Therapeutic applications of ribozymes and riboswitches. *Curr. Opin. Pharmacol.*, **10**, 551–556.

Thiel, K.W. and Giangrande, P.H. (2009) Therapeutic applications of DNA and RNA aptamers. *Oligonucleotides*, **19**, 209–222.

Watts, J.K. and Corey, D.R. (2012) Silencing disease genes in the laboratory and the clinic. *J. Pathol.*, **226**, 365–379.

Wang, X., Huang, X., Fang, X., Zhang, Y., and Wang W. (2016) CRISPR-Cas9 system as a versatile tool for genome engineering in human cells. *Mol. Ther. Nucleic. Acids*, **5**, e388.

Proteome Analysis

39

In 1975 O'Farrell and Klose independently published articles with spectacular images in which they showed that the combination of isoelectric focusing and SDS gel electrophoresis is able to separate extremely complex protein mixtures. This new procedure, two-dimensional gel electrophoresis, quickly became established as the most successful high-resolution technique for separating proteins. Attempts were soon made to use the information contained in the protein patterns to solve biochemical and medical questions. By comparing the protein patterns of different, defined states of a cell or a body fluid (e.g., sick or healthy, different metabolic conditions, etc.) changes in protein patterns become visible, which were characteristic for these conditions (Figure 39.1). This strategy for investigating biological questions is called a subtractive approach. However, the analytical methods for protein characterization (sequence analysis, amino acid analysis) were at that time not in a position to analyze such small amounts of protein as could be separated and visualized by 2D gel electrophoresis. To make matters worse, the proteins were embedded in a gel material that was incompatible with most protein chemistry techniques at the time. Therefore, the results of subtractive approaches were of mostly descriptive character. Significant changes in protein pattern could be recognized, but the identity of the proteins involved remained unknown. This changed only when the protein-chemical methods improved and became much more sensitive. At the same time methods were developed to cleave the proteins within the gel matrix and to extract the resulting peptides from the gel matrix. Alternative methods were developed in which proteins are transferred from the gel matrix to chemically inert membranes, where the immobilized proteins can then be directly sequenced or be further cleaved with enzymes. Triggered by the successes of this subtractive strategy in conjunction with the enhanced analytical techniques the idea was born to represent the protein pattern of a whole cell and interpret it quantitatively – this is the objective of proteome analysis.

The term proteome was introduced in 1995 by the Australian Marc Wilkins, as the so-called "protein equivalent of a genome." This term describes the complete protein pattern of a cell, an organism, or a body fluid.

Proteome The quantitative representation of the total protein expression pattern of a cell, an organism, or a body fluid under precisely defined conditions.

39.1 General Aspects in Proteome Analysis

Friedrich Lottspeich

Peter-Dörfler-Straße 4a, 82131 Stockdorf, Germany

Although the terms genome and proteome sound very similar they describe two fundamentally different things: a genome is a static entity, which is precisely defined by the type, order, and number of its nucleotides. However, in cell life, at no point in time are all the genes turned on and translated into proteins. A proteome is therefore a tremendously dynamic object that is affected by a large number of parameters (Figure 39.2). The delicate balance between protein

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

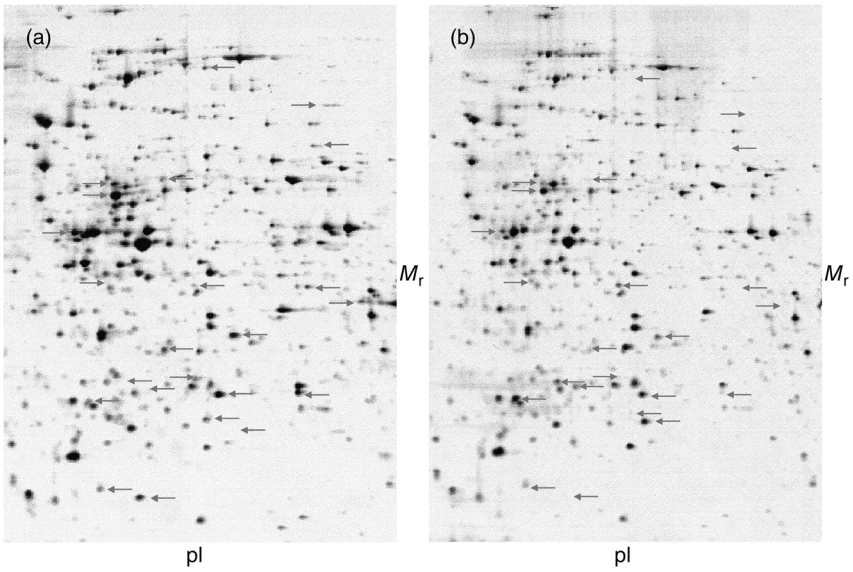


Figure 39.1 Subtractive approach: A cell (*Escherichia coli*) is brought from an initial state (a) by an event (e.g., other culture conditions) to another state (b). The protein patterns of both states are compared. The changes in protein patterns (indicated by gray arrows) can be attributed directly or indirectly to the triggering event.

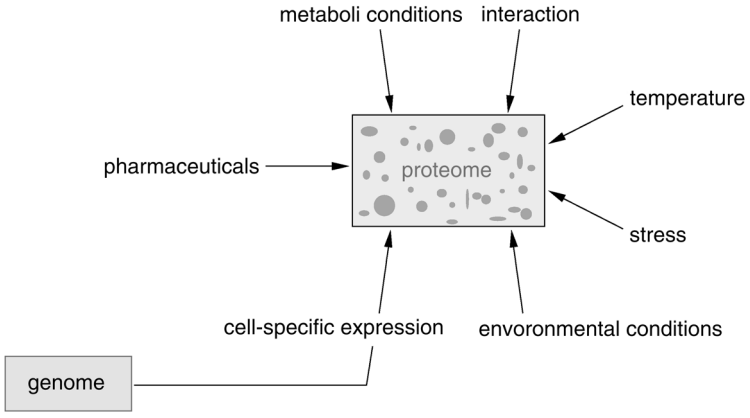


Figure 39.2 Effect of different parameters on the protein expression. The current amount of a protein in a cell is determined by various factors and is highly sensitive to changes in these variables.

synthesis, protein degradation, and many highly dynamic protein modification events can be very different under different metabolic or environmental conditions. However, the sensitive dependence of the proteome pattern on various parameters also provides the possibility to use the protein pattern as a sensitive sensor and to detect specific network-like connections by small changes in the parameters.

The great technical challenge during proteome analysis is not to change the quantitative ratios of proteins as they are present in nature. This is especially true for all steps of sample preparation and protein isolation.

Ideally, the analysis of a proteome provides the known amount of each protein. This is data that in principle cannot be obtained using molecular biology techniques, since there is no strict correlation between the amount of mRNA and the amount of the corresponding protein. Translational regulation, mRNA stability, protein stability, and protein degradation cannot be detected at the mRNA level. Therefore, it is not possible to predict from the nucleic acid level the amount of protein existing in a cell.

Clearly, from these remarks, only in rare cases is a non-quantitative proteome analysis useful. A simple determination of the protein composition of a proteome is performed only in very simple proteomes, subproteomes of organelles, or large protein complexes. In addition, this information is also usually only an initial basis for further work.

Analysis of Post-translational Modifications, Chapter 25

Another important piece of information that in principle cannot be obtained from nucleic acid sequence data is post-translational events. The active form of a protein almost always differs from the form that is present after the first translation of the mRNA. Very often the newly synthesized protein is processed, that is, amino acid residues or peptides from the N- or

C-terminal end of the translated protein are cleaved off. This is very common for enzymes that often were converted by cleavage of a peptide bond from an inactive into an active form or an active enzyme may be inactivated by processing. This mechanism of the so-called limited proteolysis plays a very important role in nature (e.g., blood clotting) and is used for the regulation of entire reaction cascades.

Other post-translational modifications take place frequently. The most common among the more than 150 known post-translational modifications are phosphorylation, sulfation, acetylation, methylation, and the linkage of certain amino acids with lipids or glycans. Almost none of the post-translational modifications is determined or can be predicted at the DNA level; but they influence decisively the biological function of a protein.

The proteome provides – contrary to the analysis of DNA or RNA – information on the quantity and the post-translational modifications of each protein.

In addition to the quantity of the expressed proteins and their post-translational modifications, the location and the neighborhood of the individual proteins is also of great importance for their interaction and thus for the function. These aspects, which are often also expected directly from the proteome analysis, are treated separately in Chapters 41 and 43.

The main objective of proteomics is to unravel network-like complex functional relationships that otherwise are extremely difficult to access. This is generally achieved by monitoring quantitatively the changes in the protein (proteome-) pattern when applying a disturbance to a given biological state (perturbation analysis). All proteome analyses have in common four complex and critical areas:

- definition of the initial conditions and the question,
- sample preparation,
- quantitative analysis of proteins,
- bioinformatics analysis.

39.2 Definition of Starting Conditions and Project Planning

If one considers that a proteome reacts extremely sensitively to any changes and is defined to a great extent by the environmental conditions (Figure 39.2) it soon becomes clear that a detailed description of all conceivable parameters has an extremely high priority. Indeed, the biological context and the environmental parameters under which a proteome analysis is applied are crucial aspects.

Several issues must be considered before a proteome analysis is performed. Only the most important are mentioned here:

- The individual proteome stages must be clearly defined, which can only be achieved by taking into account a suitable (at least reproducible) sample preparation.
- It makes sense to keep the differences between the states small to keep the number of changes in the protein pattern manageable.
- The influence and the degree of genetic heterogeneity, polymorphisms, or mutations should be kept in mind in the evaluation of differential protein patterns.
- The dynamics of proteins in living systems must be considered: biological processes at the protein level are sometimes very rapid. Regulatory processes (such as phosphorylation, dephosphorylation, transport processes, degradation processes, etc.) often take place within seconds or minutes. This means that the composition of the protein network in these short periods changes significantly and must be analyzed accordingly – an extremely challenging task, especially for project planning and preparation of samples.

  An important point in order to assess properly the expected results when planning a proteome project is the amount of sample available. Almost all protein-chemical analysis techniques work routinely on the femtomole level. For clarification: $1\,\text{mole} = 6 \times 10^{23}$

molecules; 10 fmol = 10^{-14} mol = 6×10^9 molecules. Thus, to be able to successfully analyze a relatively rare protein (1000 copies per cell), one must start at least with 10^6 cells, which is a significant amount. For extremely rare proteins that are expressed in a few copies per cell, very large amounts of starting material (10^8–10^9 cells) must be available, which is often a problem in practice. Any improvement in the sensitivity of analytical methods is therefore of utmost importance, will lead to greater depth of analysis, and will enable a more comprehensive proteomic analysis.

- In some cases, certain classes of proteins can be selectively enriched during sample preparation via affinity chromatography, special modified magnetic beads, or sample preparation arrays or chips. This has the advantage that the complexity of the original sample is significantly reduced and the following analysis is thus considerably facilitated. However, it must be carefully considered during the project planning whether the "subproteome" also contains the relevant information.
- Finally, one has to consider the statistical significance of the results: in proteomic analyses the experimental and biological variability for each set of samples has to be determined and included in the project planning and evaluation of results.

For a successful and meaningful protein-chemical determination of a protein (quantification, identification) today more than 10^9 molecules are necessary.

39.3 Sample Preparation for Proteome Analysis

The first step for quantitative proteome analysis immediately after sampling must normally be sample preparation, which differs significantly from the classical sample preparation procedures. The main objective of proteome analysis is to determine quantitative ratios of proteins in different proteomics states. In this respect, the usual steps of protein purification are inappropriate for proteome analyses. With classical methods, usually a specific protein (or a few proteins) is isolated from a complex matrix, with the focus on high purity and a good quantitative yield of the target protein(s). The other "uninteresting" proteins are treated as unimportant and are ignored. In the classical method of sample preparation and purification, typically multilevel techniques are used, all of which are inevitably connected with protein-specific losses. In addition, the separation of the proteins into individual fractions is by no means complete: proteins are mostly found in several fractions, so that the total amount of these proteins is extremely difficult to determine quantitatively.

During sample preparation for proteome analysis ideally all the proteins of a proteome must be brought into solution for subsequent quantitative analysis. The sample preparation may help to prepare a well-defined and thus meaningful proteome sample. There are virtually no limits to creative strategies, such as:

- for example, cell sorting may be important to ensure that the cells examined are all in the same cell cycle stage;
- with tumor tissue studies tumor cells could be enriched by laser microdissection, after which the cells are contaminated only with minor amounts of other cells;
- centrifugation or free-flow electrophoresis can be used for organelle separations to obtain preparations that are as homogeneous as possible.

Artificial changes in the protein composition by proteolysis or other modifications (e.g., oxidations) must be avoided. Therefore, in proteome analysis the length and reproducibility of sample preparation play an important role. Each manipulation with proteins leads inevitably to losses, mainly due to absorption to any kind of surface as a result of the hydrophobic character of proteins. Unfortunately, different proteins exhibit different degree of losses, and these are not predictable. Consequently, after the sample preparation steps the original quantitative composition of the sample is no longer guaranteed. It follows that in general the sample preparation for proteome analysis should consist of very few steps. For the next steps – the quantitative determination of the amounts of protein – different ways have emerged in recent years, which are dealt with in the following sections. Together, all these strategies have significant limitations,

which are based mainly on two properties of a proteome: the enormous complexity and the large dynamic range of proteins, which makes determining low abundance components so difficult:

- *Complexity:* For example, the human genome has probably only about 20 000 genes, but it produces by various processes on the way from the gene to the protein (e.g., by alternative splicing, mRNA editing, processing, post-translational modifications, etc.) hundreds of thousands, indeed probably even millions, of protein species (also called proteoforms). Although these different protein species may be derived from a single gene, they are at molecular level – and often functionally – clearly different. For a comprehensive proteome analysis each of these proteoforms must be characterized and quantified individually. But this causes serious and fundamental difficulties because, to date, there are no techniques available that can unravel hundreds of thousands of components. Therefore, a concept called "reduction of complexity" is often adopted, which does not, as originally intended, analyze the pattern of all proteins (proteome) but only a subset of the proteome, a subproteome. Such focusing on a subproteome is usually also associated with a restriction of the biological question that can be addressed. A typical example is the study of all phosphorylated proteins, the phosphoproteome, which is particularly important, for example, in signal transduction processes. Along the lines of this reduction of complexity is the analysis a subproteome with a targeted proteomics strategy (Section 39.5), which has proved extremely successful in the study of molecular machines (e.g., spliceosome, the ribosome, proteasome, etc.) and in the elucidation of protein complexes, which are responsible for the execution of many biological functions. Despite the successful application of such strategies, one should be aware that an essential aspect of a holistic analysis, that is, to elucidate unexpected, far-reaching, or transient functional contexts without prior knowledge, is hardly possible.
- *Low abundance proteins:* A particular challenge in a proteome analysis is the very different amounts of protein species. There are, for example, in human blood plasma proteins such as albumin that occur in the concentration range of milligrams per milliliter. But there are also important proteins – for example, the prostate-specific antigen – that occur in concentrations of picograms per milliliter (Figure 39.3). This dynamic range of protein amounts, covering ten to twelve orders of magnitude means that for a proteomic analysis we have to detect, identify, and quantify one protein molecule in the presence of one billion molecules of another protein species – an impossible task so far. Most analytical techniques can cover a dynamic range of 10^2–10^3, which is very far from the required minimal 10^8. The only way to cover a large dynamic range is to separate the frequent proteins from the rare ones and analyze them separately. A further problem is that a small number of highly expressed proteins accounts for the major part of the sample material and many regulatory or diagnostically interesting proteins occur only in very small amounts. For example, in plasma the 22 most abundant proteins account for about 99% of the total protein mass of plasma. Therefore, sometimes in plasma proteome analyses the few most abundant proteins, which account for 90% of the amount of protein in blood plasma, are removed via affinity methods

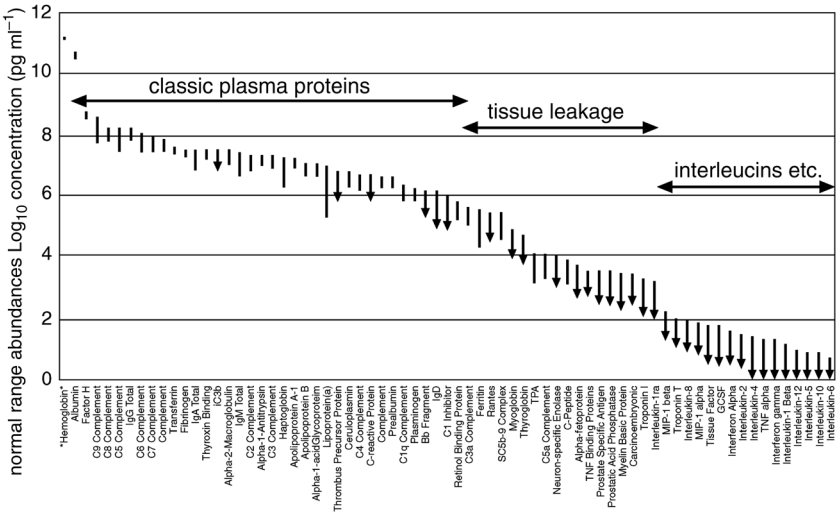


Figure 39.3 The dynamic range of the frequency of plasma proteins covers more than ten orders of magnitude. Albumin is present in the blood plasma at a concentration of 50 mg ml^{-1}, while interleukins are present in concentrations as low as a few picograms per milliliter. Source: Anderson, N.L. et. al. (2002) *Mol. Cell. Proteomics*, **1**, 845–867. With permission, Copyright © 2002, by the American Society for Biochemistry and Molecular Biology.

("depletion"). The proteomic analysis of the remaining material is then carried out with a more uniform sample. However, the completeness of the depletion is critical and is very difficult to control since even 0.1% of a highly abundant protein still appears in the depleted sample as an abundant protein and often varies quantitatively quite strongly due to the depletion process. Additionally, there is also always a risk that specific rare proteins may bind to the abundant proteins and are co-depleted. In such cases, the quantitative ratios of the depleted sample may no longer reflect those of the original sample.

39.4 Protein Based Quantitative Proteome Analysis (Top-Down Proteomics)

Nicole A. Haverland, Owen S. Skinner, and Neil L. Kelleher

Northwestern University, Departments of Chemistry, Molecular Biosciences, and the Feinberg School of Medicine, 2145 Sheridan Road, Evanston, IL 60208-3113, USA

Quantitative proteomics is today carried out mainly by two strategies. The top-down strategy has the intention to keep the proteins intact as long as possible. Reduction of complexity, quantification, and mass spectrometric identification of the individual proteins is achieved at the protein level. A typical example of a top-down approach is the classic gel-based proteomics (Section 39.4.1), when 2D gel electrophoresis is used to separate the proteins with subsequent image-based quantification. Mass spectrometry is here exclusively used for identifying the proteins. The difficulties and limitations of separation methods for proteins led to alternative or complementary techniques for quantitative proteomics, the bottom-up proteomics strategies (Section 39.5). Here, as a first step, all the proteins of a proteome are cleaved into peptides. The resulting immensely complex peptide mixture is fractionated using different separation methods and quantitatively analyzed by mass spectrometry. The peptides are identified and allocated via informatics to individual proteins.

39.4.1 Two-Dimensional-Gel-Based Proteomics

Most top-down proteome data that can be found today in the literature have been obtained by means of 2D gel electrophoresis. This strategy can be divided into the following sub-steps:

- sample preparation,
- separation of proteins,
- image analysis,
- quantification of the proteins and data analysis,
- identification and characterization of the proteins.

Sample Preparation In the first step, the proteins of the different proteome states for 2D gel electrophoresis have to be completely dissolved. To avoid interference of isoelectric focusing, no salts should be present and only zwitterionic or non-ionic detergents are used in the sample buffer. In practice, many cells can be solved directly in the buffer for the 2D gel electrophoresis (8 M urea, 2% CHAPS). For difficult samples (e.g., tissue or poorly soluble cells, such as fibrous cells) special sample preparation protocols have been prepared (French press, ultra-turrax, bead beater, etc.). Again, before separation by 2D gel electrophoresis any undissolved material should be removed from the sample by high speed centrifugation.

Separation of Proteins As described in Chapter 11, 2D gel electrophoresis is a classic separation technology that is able to provide a separation space for up to 10 000 protein species, which corresponds almost to the total number of proteins in simple cells.

Two-dimensional gel electrophoresis has many features that makes it particularly suitable for proteomics:

- high resolution, up to 10 000 components can be separated in a single gel;
- IEF pH ranges can be spread by using immobilized pH gradients (Immobilines) in the first dimension, which enables higher resolution in specific pH ranges;

- it is compatible with detergents and therefore universally applicable for all proteins; in addition, hydrophobic proteins like membrane proteins can be separated;
- using new application techniques milligram quantities of proteins can be separated in a 2D semi-preparative gel:
- two-dimensional electrophoresis is relatively fast to carry out (1–2 days).

Despite these indisputable advantages over other methods there are also some serious limitations of 2D gel electrophoresis:

- lack of automation, therefore limited reproducibility (regarding the position of the proteins in the gel as well as the amounts of individual proteins);
- difficult technical implementation;
- transfer from the first (IEF) to the second dimension (SDS-PAGE) is not complete and involves a risk of protein loss;
- the matrix is not inert, the proteins must be retrieved from the gel matrix for further analysis;
- low dynamic range capacity, that is, proteins with low copy number cannot be displayed simultaneously to proteins with high expression rate;
- only a few good methods to quantitate proteins in the gel matrix are available.

The last two points are for a proteome analysis certainly the most problematic. For proteomics, the main goal is to represent all the proteins. The low dynamic range of gel electrophoresis of a maximum of 10^3 is a serious limitation. Currently, one only can try to separate common and rare proteins as far as possible from each other (e.g., with spread pH gradient and/or by prefractionation) and to quantify them in separate analyses.

Here should be clearly pointed out that, although the gel has an extremely high separation capacity, a single spot in no way means that there is only one protein in that single spot. Even a purely statistical calculation assuming about 30 000 proteins in a more complex cellular proteome and a separation capacity of 10 000 concludes that on average three protein components in each spot must be present. This means that, most likely, in each spot in a 2D gel several proteins are present, although for analytical purposes (dynamic range, sensitivity, etc.) perhaps only one can be identified.

The quantification of proteins is a central concern of proteome analysis. The proteins need to be stained for quantification in the gel matrix. Staining characteristics are different for each dye and can be very different from protein to protein and, unfortunately, for different amounts of protein. Consequently, in principle, very small amounts of protein can adsorb more dye relative to large amounts of protein. Furthermore, small variations in the staining protocols may lead to different staining intensity.

The most popular techniques for the visualization of proteins include staining with Coomassie Blue, with a rather low sensitivity and dynamic range (detection limit about 100 ng, dynamic range about 10^2). In addition, silver staining (detection limit 10 ng) is often used, which, however, hardly leads to correct quantitative values. More sensitive are fluorescence staining, autoradiography (the detection of radiolabeled proteins), and immunological staining (if characterized antibodies are available). Owing to the limited reproducibility and the problems with protein staining, multiple determinations (5–10 gels of the same sample, preferably by independent workups) must always be performed to obtain statistically meaningful, quantitative information. Today, with technical replicates standard deviations below 15% can usually be reached for many proteins in 2D electrophoresis. Nevertheless, there are specific proteins that show much larger variations even under optimal instrumental and operating conditions.

Imaging and Quantification of Proteins, Data Analysis After separation the stained proteins must be quantified. This is done with 2D gels by densitometry with a laser densitometer or a scanner. Commercially offered software evaluates the images so that in a first step the outlines of the protein spots are detected. One may influence - depending on the gel quality - the results to some extent through changing the input parameters. Normally not all spots are identified correctly even with an optimal set of parameters. However, a major limitation is that an error of even a few percent on average will with 2000 spots result in 40–100 incorrectly captured proteins. Even though software support has improved markedly in recent years,

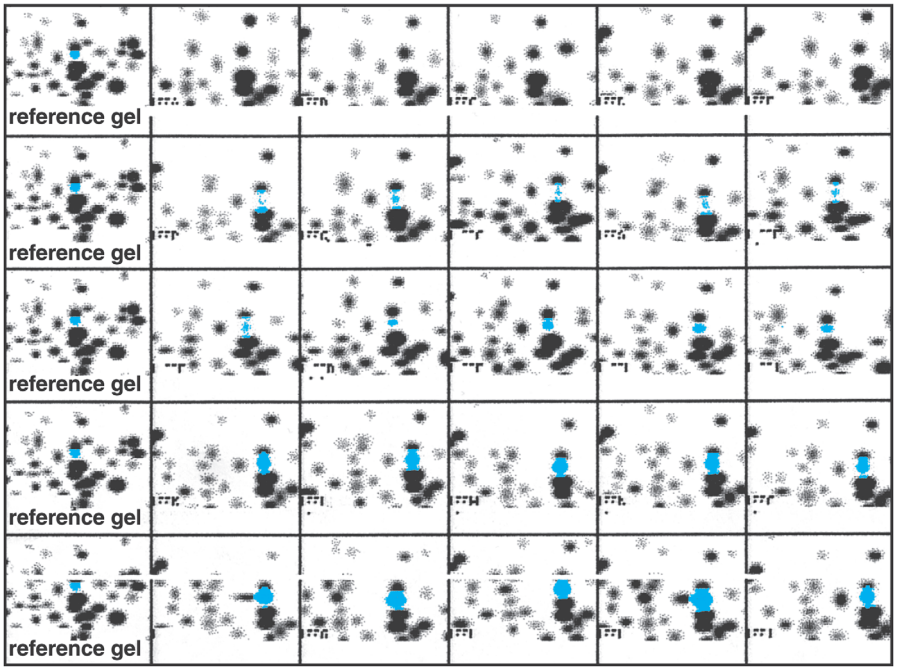


Figure 39.4 Excerpts of 25 gels of different experiments, which by computer-assistance are compared with reference gels. The gels are arranged in groups according to the expression of a particular protein (blue). One can see the reproducibility of the protein pattern and the significant differences in the expression of the tagged protein.

significant, time-consuming editing work has to be carried out at this point. After defining the protein spots they are automatically quantified relative to each other and the results stored in databases.

Powerful software tools make it possible to compare gels, to correct small distortions, and to represent differences in the protein patterns of different gels. A form of presentation is partly reproduced based on a comparison of 25 two-dimensional gels with controls in Figure 39.4. Even at this level differential protein patterns and significant differences in the quantitative ratios can be recognized and analyzed. Above all, statistical data must be obtained and averaged statistically significant reference gels of the individual states must be prepared.

Identification and characterization protocols of the proteins for further analytical procedures are optimized for high throughput and mainly use mass spectrometry.

Proteins present after the separation in a 2D gel may be analyzed mainly in two ways. One possibility is that the intact protein is transferred from the gel matrix onto a chemically inert membrane and protein chemical analyses (e.g., sequence analysis) are performed directly on the membrane. Alternatively, the protein is enzymatically cleaved into smaller fragments directly in the gel matrix, and the resulting peptides are extracted and analyzed. Both methods complement each other and often give very specific identifications (Figure 39.5).

For analysis of the intact protein in a first step, all proteins are electro-blotted from the gel matrix to a PVDF membrane. The transfer may be almost quantitative for average proteins, while for large or hydrophobic proteins rather low transfer yields can be expected. Thus,

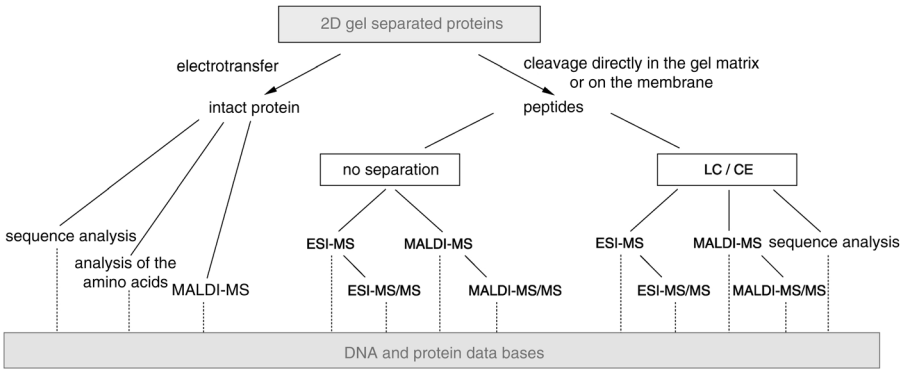


Figure 39.5 Analysis of proteins separated by gel electrophoresis. After separation in the 2D gel, a protein may be analyzed either directly by electrotransfer to a membrane or after enzymatic cleavage, on the basis of the resulting peptide fragments.

significant losses may occur and the quantification after electrotransfer certainly no longer reflects the original conditions.

A direct amino acid sequence analysis on the membrane may be successfully performed. The minimum amount required for sequence analysis is currently in the low picomole range. As a general rule a visible, Coomassie stained spot analysis in the gel/on the blot is sufficient for sequence analysis. By amino acid sequencing a protein will often be identified by a protein database search. Unfortunately, many of the naturally occurring proteins are N-terminally blocked and will not give any result.

The analysis of total proteins for identification and characterization by Edman amino acid sequence analysis (Chapter 14) has the disadvantage that it takes a relatively long time, and is also complex and can only provide limited information. Therefore, today, for proteome analyses the characterization and identification of mainly tryptic fragments by mass spectrometry are optimized for high-throughput analysis.

Mass Spectrometry, Chapter 15

Analysis of Peptide Fragments For the analysis of peptides the proteins must be cleaved enzymatically. This can take place directly in the gel matrix or after transfer of the proteins on an inert membrane. However, since the membrane transfer is an additional step that can be fraught with losses the cleavage is preferably carried out directly in the gel matrix. The enzymes that are used for this are in particular trypsin, endoprotease LysC, and endoprotease AspN. After cleavage, the resulting peptide fragments are eluted from the gel matrix using mostly organic solvents and acids. It is not expected that all peptides from the gel matrix can be completely recovered – especially hydrophobic or large peptides often provide poor yields or are not eluted at all. The extracted peptides must be further analyzed, for which in particular mass spectrometry techniques are applied. The individual steps are shown in the overview in Figure 39.5.

Cleavage of Proteins, Chapter 9

For mass spectrometric identification the peptide mixture, which was previously desalted through a reversed phase mini-cartridge, is analyzed either with MALDI-MS or ESI (nano-spray-) MS. In both cases, the sample output consists of a list of the mass-to-charge states of all peptides identified. The measured values are then compared with a list of peptide masses that were generated *in silico* by theoretical cleavage of all proteins of an organism. Ideally, all the measured peptide masses should be found in the theoretical peptide mass list of a protein. In practice, however, only less than 30–70% of the measured values can be directly assign to a single protein, but virtually always sufficient to clearly identify the protein. The remaining mass values can sometimes stem from oxidation products, unexpected fragmentations, modifications, or even from peptides of contaminating proteins.

An alternative to analysis of the peptide pattern is the mass spectrometric sequencing of the proteolytic peptides. It provides a more confidence of identification. With mass spectrometric fragmentation techniques (using MALDI-TOF/TOF or ESI-MS/MS instruments) tandem MS spectra can be generated. Using special search algorithms, the measured spectra are compared with the database of computer generated theoretical MS/MS (= MS^2) spectra of all peptides of the same organism.

The automatic interpretation of MS/MS spectra is increasingly being carried out also for the *de novo* sequence analysis of unknown peptides, wherein the limiting step is the interpretation of the fragment spectra. If the identification of mass spectrometric analysis failed or is ambiguous, classical protein chemistry methods must be applied (Figure 39.5). Here again, of course, the mass spectrometric methods play a prominent role, but they are used in conjunction with HPLC or CE separation of peptides, with Edman sequence analysis, or with other analytical techniques. In this case, the throughput is of course much smaller than with the typical proteomics workflow.

Chromatographic Separation Techniques, Chapter 10

Capillary Electrophoresis, Chapter 12

Amino Acid Sequence Analysis, Chapter 14

When a protein is identified via peptide sequences in a (DNA) database, the protein is then further analyzed only if post-translational modifications are suspected. This may be the case if the measured molecular weight or the observed isoelectric point of the protein deviates from the theoretically predicted values. Significant deviations (more than about 0.3 pH units) are most probably due to one or more post-translational modifications. If very accurate information is required – such as in the characterization of recombinant therapeutic proteins – also without evidence of a modification each amino acid residue has to be covered by an experimental analytical data value. Most often then the protein must be cleaved in independent experiments with different enzymes. Even today these analyses, whereby any modification and even the

Sequence Data Analysis, Chapter 33

Analysis of Post-translational Modifications, Chapter 25

position of a modification has to be determined exactly, are extremely challenging, time consuming, and often the whole arsenal of classical protein chemistry, including mass spectrometric techniques, must be used.

39.4.2 Two-Dimensional Differential Gel Electrophoresis (2D DIGE)

Two-dimensional DIGE is a variation of the 2D gel electrophoresis, which offers some significant advantages over the classical technique. Two (or three) protein extracts are labeled individually with different fluorescent reagents (Cy-dyes with reactive *N*-hydroxysuccinimide). The reagents bind covalently to the ϵ-amino group of lysine residues in the protein. A positive charge compensates for the loss of the positive charge of the lysine residue due to the derivatization. In addition, the masses of the individual reagents are rather small in order to avoid a different migration behavior between the labeled and the unlabeled protein. The mass and electrophoretic properties of the different modification dyes are almost identical to assure that the migration behavior in a 2D gel is almost identical for the corresponding protein labeled with the different fluorescence dyes. There are two versions of the DIGE technique in use. In "minimal labeling" the reaction conditions are chosen so that only a small percentage of the lysine residues of a protein molecule is marked. In "maximum labeling" the cysteine residues of the proteins are fully derivatized with the dye. The reason for using the cysteine modification is that cysteine is a rather rare amino acid and derivatization with the hydrophobic reagent (dye) does not lead to insoluble proteins. The complete modification of a frequently occurring amino acid (e.g., lysine) would often lead to hydrophobic and insoluble proteins.

DIGE, Section 11.6.6

The individual fluorescent dyes have different spectra. Therefore, one can combine the individual differently labeled protein extracts and separate them into a single, conventional 2D gel. The proteins of the individual samples can be individually visualized by image recording with the appropriate excitation and emission wavelength filters. Quantitative comparison of protein patterns is carried out by using special software programs.

With this technique, the serious gel-to-gel variations of the classic 2D gel electrophoresis can be avoided. This allows comparative, quantitative proteomics with fewer gels, less material consumption, higher accuracy, and in less time. The difficult spot matching of different gels is eliminated because the entire analysis is carried out in a single gel. The fluorescent dyes used exhibit very high detection sensitivity and a higher dynamic range by covering more than five orders of magnitude in a linear calibration curve.

Since the protein spots are only visible under UV light, an automatic spot picker is often necessary to transfer the protein spots to a mass spectrometric identification. As more proteins can migrate in the same spot and the small amount of fluorescent protein migrates at a slightly higher mass than the unmarked protein, there is a risk that the picked protein spot (identified by MS) is not identical to the protein identified and quantitated in the imaging process.

As with all proteomic techniques, with 2D DIGE the statistical experimental error also has to be determined and taken into account when interpreting the results. The experimental error in 2D DIGE arises primarily in the sample preparation, variation in the protein labeling, and errors in image processing (bad spot detection and background problems due to different fluorescence characteristics of the acrylamide at different wavelengths).

39.4.3 Top-Down Proteomics using Isotope Labels

The main challenge in protein-based proteomics is to preserve the quantitative ratio of the individual protein species in different proteome states. However, proteins as rather large molecules are sensitive on all levels of structure (primary, secondary, tertiary, and quaternary structure). Small changes/damages or a different environmental parameter (which is very common when comparing two or more different proteome states over a long time) may lead to differential behavior of certain proteins during sample work up and fractionation. Interactions with surfaces, chromatographic column material, gels in electrophoresis, or with other proteins will cause loss of proteins, unfortunately in an entirely unpredictable manner. Therefore, it is rather difficult in a top-down proteomics approach, when comparing two or more different

samples and performing independent steps to fractionate and reduce complexity, to keep the quantitative ratio of proteins unchanged. On the other hand, for high quality results this reduction of complexity is a prerequisite for almost any analysis technique due to the diversity and complexity of a proteome. Independent direct proteome analyses from the same sample – even without sample preparation or fractionation steps – at present do not lead to exactly the same output.

Introducing the concept of isotopic labeling and multiplexing may overcome several of the problems associated with independent multiple label-free analyses. A detailed evaluation of the concept and availability of stable isotope methods for top-down proteomics is given in Section 39.6.1.

39.4.4 Top-Down Proteomics using Intact Protein Mass Spectrometry

Top-down proteomics refers to the comprehensive analysis of intact proteins using mass spectrometry. Whereas bottom-up proteomics relies on peptides derived from enzymatic digestion of proteins, top-down proteomics utilizes the intact protein for analysis, including its modifications and other forms of sequence variation. The advantage of this approach is the ability to detect and measure single-nucleotide polymorphisms, isoforms, splice variants, post-translationally modified proteins, and even whole protein complexes. Unsurprisingly, everything changes when enzymatic digestion is excluded. In particular, top-down proteomics avoids the imperfect process of rebuilding a set of related protein forms using a collection of peptides.

Throughout this chapter, the basic concepts and terminology for intact protein mass spectrometry is covered, measuring intact mass (MS^1), fragmentation and the measurement of product ions (MS^2), data analysis, and high-throughput top-down proteomics. To begin these discussions, a classic example of top-down mass spectrometry analysis is given: the 76 amino acid protein, ubiquitin. Although there are a multitude of approaches available for intact protein mass spectrometry, the discussions here will center on high-resolution data that was acquired using electrospray ionization and a Fourier-transform mass spectrometer operating in positive ion mode.

Mass Spectrometry, Chapter 15

39.4.5 Concepts in Intact Protein Mass Spectrometry

At first glance, it is easy to notice the differences in the mass spectrum of an intact protein (Figure 39.6a) as compared to that of a peptide. The set of major peaks that are shown in Figure 39.6 actually represent a single protein, ubiquitin. This is in contrast to a peptide MS^1 spectrum, in which distinct peptides are typically observed as 1+, 2+, and/or 3+ ions. This set of major peaks in the protein MS^1 represents the different charge states of ubiquitin that were desolvated and transmitted into the gas phase using electrospray ionization. Although the interpretation of an intact protein mass spectrum may seem intimidating, the basic concepts presented in this chapter will provide the framework needed for analyzing and understanding top-down proteomics data.

In Figure 39.6b, an enlargement of the mass spectrum shows that the 11+ charge state peak is not a single signal but rather consists of a distribution of peaks. Known as isotopomers, each of these peaks has the same chemical formula and charge but differs in the number of heavy isotopes present in the molecule. The multitude of different masses from the same protein results in two concepts that are important to understand: the monoisotopic mass and the average mass. The monoisotopic mass is calculated using the exact mass of the most abundant isotope for each element present in the molecule. In small peptides, such as the one shown in Figure 39.7a, the monoisotopic mass is the most abundant mass in the spectrum. However, in larger proteins (Figure 39.7b), the monoisotopic mass can have a very low abundance. This phenomenon can be attributed to the increased likelihood of at least one of each element being a heavy isotope as the overall number of elements increases. This leads to the average mass, which refers to the weighted average of the atomic masses for each isotope for each element that is present in the molecule. The average mass for the peptide and protein are indicated in Figure 39.7. In proteins, the relative abundance of each isotopomer is influenced by all of the

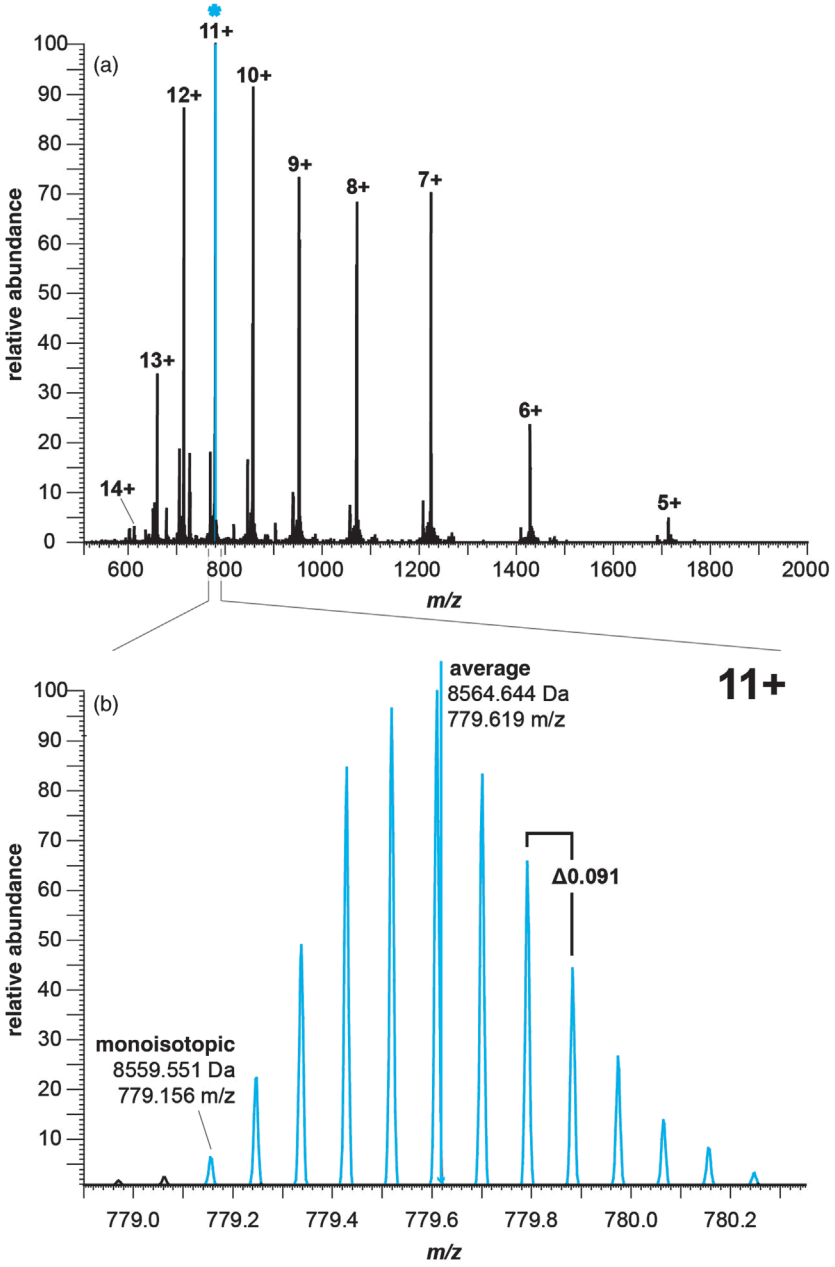


Figure 39.6 A top-down mass spectrum of ubiquitin. (a) The charge state distribution of the intact protein and (b) the isotopic distribution of the 11+ charge state with the monoisotopic and average mass indicated.

elements that make up the molecule. However, because carbon has the most abundant naturally-occurring heavy isotope it has the largest effect on the overall isotopic distribution. Thus, the number of heavy isotopes in each peak is indicated by the subscript next to the ^{13}C; for example, $^{13}C_5$ represents the protein with exactly five heavy isotopes. The difference between average and monoisotopic mass can be remarkably large and increases with protein size. Table 39.1 highlights this point using examples of large, medium, and small proteins as compared to a peptide.

As previously highlighted, one of the most distinguishable differences in the mass spectrum of an intact protein as compared to that of a peptide is the presence of multiple charge states. These charges states arise as a result of solution equilibria and electrospray ionization, which for positive ion mode result in protonation of basic residues found throughout the protein. Unlike peptides, proteins often contain multiple basic residues that can be protonated, which creates a distribution of charge states. The number of accessible, ionizable residues influences the number of charge states for a given protein. As a general rule of thumb, the total number of

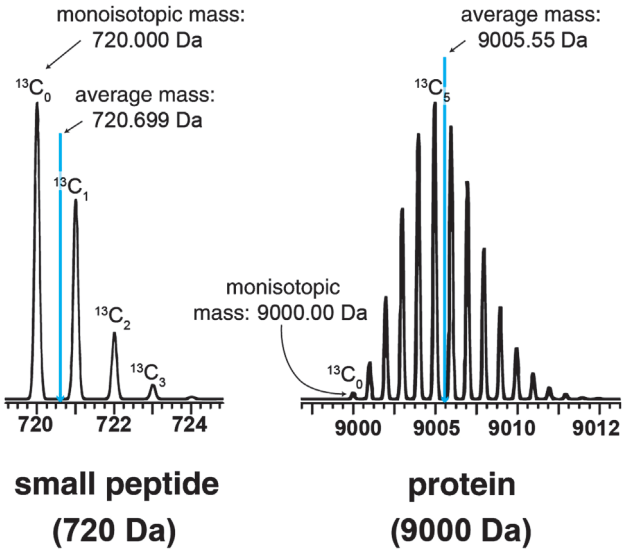


Figure 39.7 Isotopic distributions and monoisotopic and average mass. A comparison for the isotopic distributions of a small peptide (a) and a protein (b). The monoisotopic and average masses are indicated for each.

Table 39.1 The average and monoisotopic masses of four human proteins and a peptide.

Uniprot accession number	Name	Amino acid length	Monoisotopic mass (Da)	Average mass (Da)	Difference (Da)
98WZ42	Titin	34350	3813651.757	3815992.986	2341.229
P02787	Serotransferrin	679	75146.569	75194.920	48.351
P62979 [1-76]	Ubiquitin	76	8559.617	8564.757	5.140
Q8IVG9	Humanin	24	2685.482	2687.240	1.758
P01858	Phagocytosis-stimulating peptide	4	500.307	500.594	0.286

observable charge states is roughly equal to the intact mass of the protein divided by 1000. In Figure 39.6a, ten distinct charge states are observed for ubiquitin and range from 14+ to 5+. To determine the average mass or the monoisotopic neutral mass of ubiquitin using the experimental data, we employ the equation:

$$M = mz - M_H z$$

where: M is the average or monoisotopic neutral mass of the protein,
m is the observed mass-to-charge (m/z) ratio,
z is the observed charge,
M_H is the mass of a proton, which is generally considered to be 1.00727 Da.

For theoretical work, this equation can be rearranged to solve for the m/z ratio for a given charge state or set of charge states:

$$m/z = \frac{M + M_H z}{z}$$

As provided in Table 39.1, the monoisotopic mass and the average mass of ubiquitin are 8559.617 Da and 8564.757 Da, respectively. Using this equation, we can determine the theoretical monoisotopic and average m/z values for each of the ten charge states that were observed for ubiquitin (Figure 39.6a; Table 39.2).

The charge state can be directly measured using a high-resolution mass spectrometer that is capable of resolving the peaks of a protein's individual isotopomers. A collection of isotopomers is known as the isotopomer envelope and measuring the distance between isotopic peaks within an envelope can directly indicate the charge state. Because isotopes differ by one neutron, the spacing between isotopic peaks will be ~1 divided by the charge of the ion. For example, Figure 39.6b presents the isotopic distribution for the 11+ charge state of ubiquitin.

Table 39.2 Monoisotopic and average *m/z* values for ten observed charge states from ubiquitin.

Charge state	Monoisotopic *m/z*	Average *m/z*
M	**8559.617**	**8564.757**
5	1712.931	1713.959
6	1427.610	1428.467
7	1223.810	1224.544
8	1070.959	1071.602
9	952.076	952.647
10	856.969	857.483
11	779.154	779.622
12	714.309	714.737
13	659.439	659.835
14	612.408	612.776

The distance between isotopic peaks is 0.091, which is an isotopic spacing of ~1/11 *m/z*. Taken together, using high-resolution mass spectrometry, it is possible to isotopically resolve a single ion species and subsequently determine its charge state. Furthermore, the observation of a protein with isotopic resolution provides the means to calculate a much more accurate mass value, often with an error tolerance of <10 ppm; for the case of ubiquitin, the calculated monoisotopic mass from the 11+ charge state (8559.636 Da) matches to the theoretical mass (8559.617 Da) with an error of 2.3 ppm (Figure 39.6b).

Using Mass Spectrometry to Measure Intact Proteins The dominant peaks in Figure 39.6a correspond to the full-length intact mass of ubiquitin. However, closer inspection reveals a set of satellite peaks that are both lower in abundance and lower in mass than full-length ubiquitin (Figure 39.8a). To determine the neutral monoisotopic mass of both of these protein species simultaneously, one can *deconvolute* the mass spectrum to generate a monoisotopic mass spectrum and/or list (Figure 39.8b). Spectral deconvolution uses an algorithm to calculate the neutral mass of a protein species using the information provided within an isotopomer envelope, including the determination of charge state for that isotopomer envelope as previously mentioned. The relative intensity for each protein species includes contributions from all charge states observed for the neutral mass. As displayed in Figure 39.8b, the *y*-axis remains as relative abundance; however, the *x*-axis has changed from a *m/z* value to a monoisotopic neutral mass measured in Da. Furthermore, two major peaks are observed, one at 8445.585 Da and the other at 8559.621 Da. The minor peaks that are observed can represent oxidation and/or neutral loss products. The mass difference between the two major protein species is 114.035 Da. Although intact masses for each of these protein species can be acquired, without further characterization the identity of this mass shift will remain unknown.

The ability to unambiguously identify and characterize a protein requires more than just an intact neutral mass. Improved characterization will occur by *isolating* and *fragmenting* a specific ion to produce fragmentation products, or product ions, that can then be traced back to the isolated *precursor*. Fragmentation is a type of dissociation that generates predictable product ions, which are dependent on the fragmentation approach (Table 39.3).

In the case of ubiquitin, higher-energy collisional dissociation (HCD) was used to generate *b*- and *y*-type fragment ions for each of the isolated protein species; this is also known as MS^2. These fragment ions were then measured using a high-resolution mass spectrometer. The mass spectra generated during MS^2 of intact proteins resemble that of peptide MS analyses, in that the occurrence of multiple charge states from a single species is severely decreased. In Figure 39.9a, a butterfly MS^2 spectrum is provided for each of the individually isolated major protein species observed in Figure 39.8a. Although there is overlap between each of the spectra, several peaks appear to be shifted; an example of both effects is provided in Figure 39.9b. The $y58^{8+}$ and $y56^{8+}$ are displayed in purple and the 114.060 Da shift is once again observed. This is in contrast to the $b18^{2+}$ ions that are presented in orange, which exhibit no mass shift between

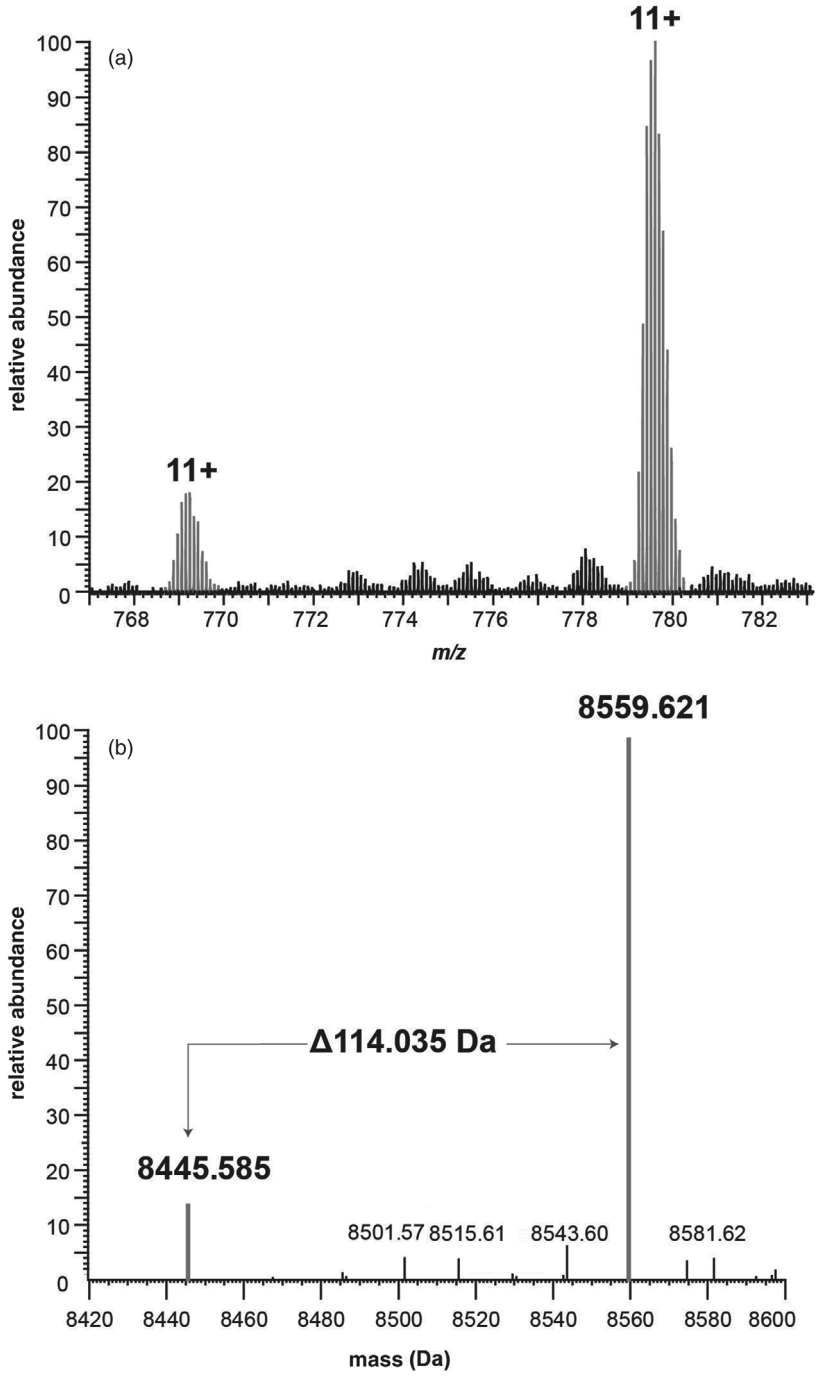


Figure 39.8 Two forms of ubiquitin. (a) A spectrum of the intact 11+ ubiquitin ions showing the isotopic distributions of the protein and a lower-mass satellite peak. (b) The deconvoluted monoisotopic masses of each of these species indicating the 114.035 Da difference in mass.

species. To better interpret this complex data, the same deconvolution approach previously described can be employed and in doing so a list of neutral fragment masses can be compiled that are generated for each of the isolated precursor ions. These neutral fragment masses can then be used for the identification and characterization of distinct protein forms, which are termed proteoforms.

Data Analysis Neutral masses obtained for fragment ions can be correlated to a theoretical sequence using a graphical fragment map. This approach streamlines data analysis and interpretation by simplifying a complex dataset (i.e., a list of neutral fragment masses) into an easy to read and interpret graphic complete with the overall level of characterization for a

Table 39.3 Gas-phase fragmentation approaches.

Type of fragmentation	Abbreviation	Fragment ions
Collisionally activated dissociation/collisionally induced dissociation	CAD/CID	b/y
Higher energy collisional dissociation	HCD	b/y
Infrared multiphoton dissociation	IRMPD	b/y
Sustained off-resonance irradiation–collisionally activated dissociation	SORI-CAD	b/y
Electron capture dissociation	ECD	c/z
Electron transfer dissociation	ETD	c/z
Electron transfer higher-energy collisional dissociation	EThcD	b/c/y/z
Ultraviolet photodissociation	UVPD	a/b/c/x/y/z

given protein. Graphical fragment maps contain information concerning the theoretical neutral fragment masses for a given sequence, which can then be used to align the experimental neutral fragment masses within a predefined error tolerance (e.g., 10 ppm). Experimentally derived fragmentation products that match the theoretical values are identified with a flag. The direction, shape, and color of the flag can be used to define the type of fragmentation, which is important in hybrid analyses, such as those that use methods with both HCD and ETD (electron transfer dissociation) fragmentation sources.

In Figure 39.10a, all of the flags represent terminal fragment ions that were generated using HCD. The direction of each flag indicates that the fragment ion was either a *b*-ion or a *y*-ion, which are a N-terminal ion or a C-terminal ion, respectively. Terminal fragment ions are

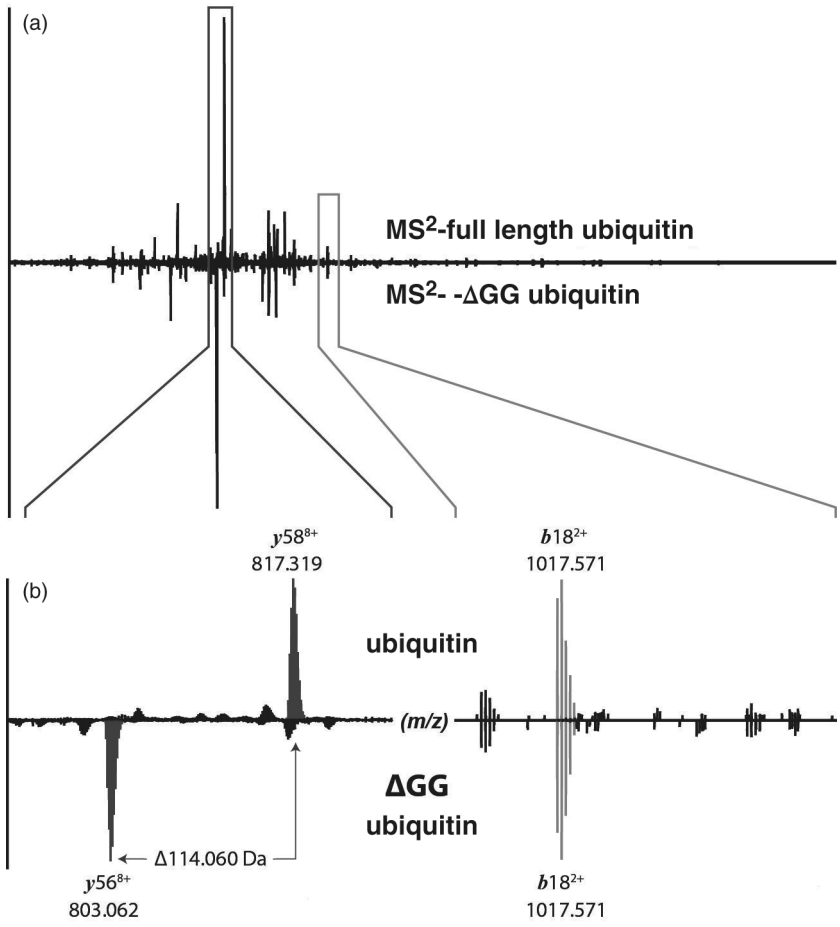


Figure 39.9 Fragmentation spectra from each of two major ubiquitin proteoforms shown in Figure 39.8. (a) The complete fragmentation spectra (MS^2) from the intact ubiquitin (top) and the ΔGG form (bottom, inverted). (b) Enlargement of two spectral features highlights the different types of fragment ions that do (left, purple) and do not (right, orange) contain the mass difference between the two proteoforms as interrogated by top-down tandem MS.

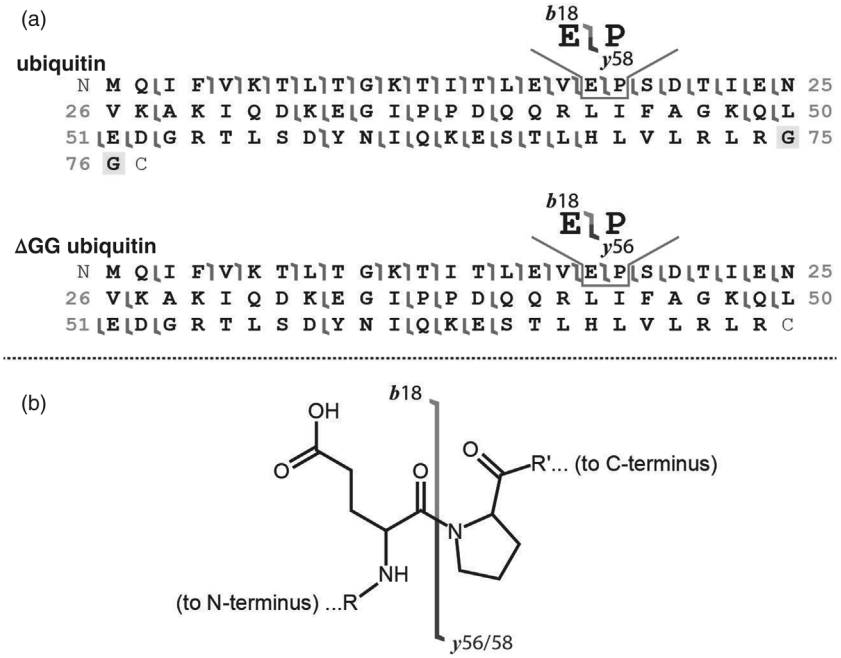


Figure 39.10 Visualizing protein fragmentation. (a) Fragment maps of ubiquitin with and without the C-terminal GG residues (yellow). The enlarged E-P cleavage indicates the fragments shown in Figure 39.8b. (b) Chemical structure showing the location of the cleavage.

product ions that contain either the N-terminus or C-terminus. A bidirectional cleavage, such as that which occurred between residues 18 and 19, indicates that both the *b*-ion ($b18^{2+}$) and complementary *y*-ion (either the $y56^{8+}$ or $y58^{8+}$) were detected. Figure 39.10b provides a visual representation of the bidirectional cleavage at the E-P cleavage site for the two ubiquitin proteoforms highlighted above in this chapter.

Further complications in data analyses can arise from the possible creation and observation of internal fragment ions. These ions contain neither terminus and are the product of both an N- and C-directional fragmentation. Unfortunately, internal fragment ions require more advanced computational approaches for data processing and accurate mapping of an internal fragment and terminal fragment. For example, as ubiquitin is a 76 amino acid protein there are, for a single charge state, a potential 148 terminal fragments with a minimum length of two amino acids. The equation used to determine the potential number of terminal fragments is:

$$\text{terminal fragments} = 2(n - l)$$

where n is the sequence length, l is the minimum terminal fragment length to be considered, and 2 is a constant representing the generation of N-terminal ions and C-terminal ions. In contrast, there are potentially 2701 distinct internal fragments that are possible for a single charge state. The equation used to determine the potential number of internal fragments is:

$$\text{internal fragments} = \frac{\left[(n - l - 1) + (n - l - 1)^2\right]}{2}$$

where:

n is the sequence length,
l is the minimum terminal fragment length to be considered, -1 is a correction for the elimination of terminal ions,
2 is a constant representing the generation of both N-terminal ions and C-terminal ions.

Considering that proteins often display multiple charge states when ionized, both the terminal and internal fragment ion numbers dramatically increase when analyzing multiple charge states.

By applying all these concepts, we can now better interpret the data presented by a graphical fragment map. Figure 39.10a displays a graphical fragment map that represents each of the isolated precursors previously described in Figure 39.8a. Notice that the number of matching neutral fragment masses is not the same for each of the isolated precursors. When analyzing

isolated protein species, an important consideration is the relative abundance of each. As expected, lower abundance proteins will often have a less complete fragmentation pattern than their higher abundance counterparts. Based on the deconvoluted MS^1 data (Figure 39.8b), the relative abundance of the lower abundance protein is approximately 15% of that of the higher abundance protein. Not unsurprisingly, there are fewer matching fragments for the lower abundance protein (36 fragments) than for the higher abundance protein (48 fragments). This difference can be accounted for when considering the decreased signal-to-noise ratio that occurs for a lower abundance protein as compared to a higher signal-to-noise ratio for the higher abundance protein (Figure 39.8).

Several interesting trends can be observed by aligning the experimental neutral fragment masses to the theoretical neutral fragment masses for each of the proteins. The *b*-ion cleavage patterns are highly similar between each of the isolated proteins (Figure 39.9). Because the *b*-ions are N-terminal ions, it can be concluded that the N-terminal region of both isolated proteins is identical. In contrast, the *y*-ions display a 114.060 Da shift when comparing both isolated proteins. Furthermore, because *y*-ions are C-terminal ions, one can anticipate that the 114.060 Da shift occurs at the C-terminus or at some location between the most C-terminal *y*-ion and the C-terminus. A mass shift of −114.060 Da does not match a common post-translational modification, but does correspond with a glycine-glycine removal (ΔGG). Thus, it can be ascertained that the high abundance proteoform observed in Figure 39.8a corresponds to full-length ubiquitin whereas the lower abundance proteoform corresponds to ΔGG ubiquitin. It is important to note the changes that occur for fragmentation site labeling. In the case of the ΔGG ubiquitin protein form, the sequence is modified at the C-terminus. Because of this modification, the equivalent cleavage sites for all N-terminal ions will remain constant between both the ΔGG and full-length protein forms. However, equivalent cleavage sites for C-terminal fragments will display a two-residue difference between the two protein forms (e.g., $y56^{8+}$ versus $y58^{8+}$; Figure 39.8a).

The ΔGG ubiquitin and full-length ubiquitin are two protein forms, or simply proteoforms as denoted above. The terminology surrounding whole protein molecules has been the subject of some discussion in whole protein mass spectrometry over the years. The term protein species derives from a more chemocentric view (endorsed by IUPAC). In contrast the term proteoform is gene-centric, with a focus on emphasizing biologically occurring forms (for additional details see Smith, L.M. and Kelleher, N.L., and The Consortium for Top Down Proteomics. (2013). Proteoform: a single term describing protein complexity. *Nat Methods*, **10**.(3), 186–187). The term proteoform was developed to describe the different molecular forms of discrete proteins by taking into account the canonical sequence, endogenous proteolysis, splice variants, coding single nucleotide polymorphisms, mutations, and post-translational modifications. Although bottom-up proteomic approaches are also capable of measuring distinct proteoforms, only whole protein mass spectrometry can provide a robust analysis of an intact protein. For example, the ΔGG and the full-length ubiquitin proteoforms only differ by the presence or absence of two glycine residues at the C-terminus. Using traditional bottom-up approaches, this modification will often not be detected. Trypsin is a commonly used protease in bottom-up proteomic approaches. As visualized in Figure 39.11, trypsin cleaves immediately before the

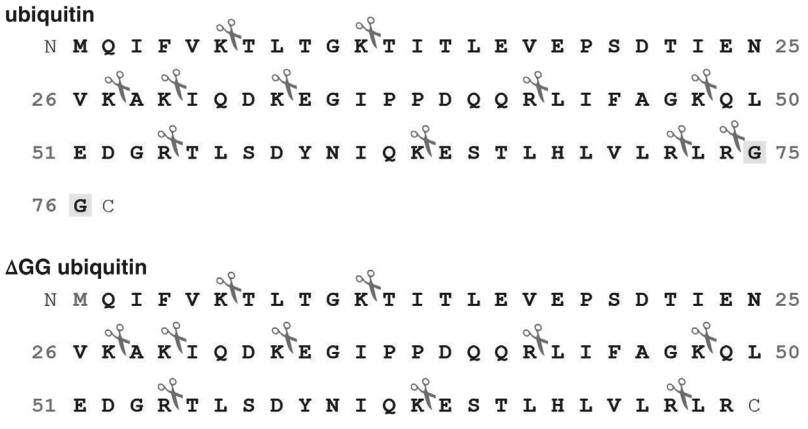


Figure 39.11 Theoretical trypsin cleavage sites in ubiquitin proteoforms. *In silico* tryptic digest shows the inability to distinguish intact ubiquitin from the ΔGG form, which is highlighted in grey.

ΔGG modification and, thus, all the measurable peptides for both proteoforms would be identical. In the case of the ubiquitin ΔGG modification, a different protease would be required to detect the modification, which will require additional considerations for downstream data processing and interpretation. Furthermore, an additional concern in bottom-up proteomics is the ability to quantify differences in the relative abundances for each proteoform and for characterizing multiple modifications. Using intact protein mass spectrometry, these limitations are largely avoided.

High-Throughput Top-Down Proteomics High-throughput top-down proteomics refers to the ability to analyze from hundreds to thousands of proteoforms in a single mass spectrometry run. Of course, to analyze these complex samples, the proteoforms must be separated into less complex mixtures in order to prevent limitations of the dynamic range. There are two types of dynamic range to consider when designing a high-throughput top-down proteomics experiment: the dynamic range of protein expression in a given system (biological dynamic range) and the instrumentation dynamic range. For example, the human plasma proteome is one of the more complex human proteomes with a dynamic range of protein expression over ten orders of magnitude. Human plasma contains high abundance serum proteins (e.g., albumin, immunoglobulins, plasminogen, etc.) as well as lower abundance interferons and clinically significant cell and tissue leakage products. Often, when analyzing human plasma, depletion of the most abundant serum proteins is performed to improve the detection of lower abundance proteins. This is because the dynamic range of instrumentation is much lower than the biological dynamic range. Furthermore, as previously described, intact proteins have multiple charge states when analyzed as an intact protein. Therefore, multiple proteins observed simultaneously will have overlapping signals (Figure 39.12a). This overlap in charge states complicates data analysis and processing as well as potentiates issues with dynamic range, such as ion masking and suppression effects brought on by high abundance proteins. Thus, the separation of complex samples prior to top-down analysis is critical to minimize dynamic range issues and charge state overlap (Figure 39.12b). Using LC separation, proteins will elute in a far less complex fashion, thereby easing computational demands for data analysis (Figure 39.12c and d). Data analysis can then proceed using the same approach as described earlier in this chapter.

As was the case with the ΔGG and the full-length ubiquitin proteoforms, it is possible to monitor distinct proteoforms simultaneously using liquid chromatography (LC) coupled to mass spectrometry (LC-MS). Of course, this ability is dependent on the type of modification. In Figure 39.8a, we observe both the ΔGG and the full-length ubiquitin proteoforms. Proteins with many post-translational modifications, such as the histone proteins, are also observed simultaneously. As shown in Figure 39.13, Histone H4 has an array of proteoforms, which are primarily a result of acetylation (+42.011 Da) and methylation (+14.016 Da) events. Isolating and fragmenting each precursor provides an opportunity to map each post-translational modification event to a specific location. Using this approach, one can determine that all ten identified histone forms contain an N-terminal acetyl moiety. The proteoform that is represented in peak 1 is only N-terminally acetylated. The mass shift from peak 1 to peak 2 is +14 Da, which represents an additional methylation. The location of each post-translational modification for each of these histone H4 proteoforms has been fully characterized and mapped, as presented by Pesavento *et al.* (Pesavento, J.J. *et al.* (2008) *J. Biol. Chem.*, **283** (22), 14927–14937). Interfering substances, such as sodium ions, can bind to the protein and can mask lower abundance peaks. As indicated in Figure 39.13, sodiation events induce a 21.983 Da mass shift by a single sodium cation replacing a single proton in a given ionized species of a proteoform (sodium mass: 22.990 Da; mass of a proton: 1.007 Da; mass shift: 22.990 − 1.007 = 21.983 Da).

Whereas proteoforms with small mass shifts often elute simultaneously in LC-MS experiments, larger modifications (e.g., splice variants, proteolytic processing) become more difficult to monitor and quantify simultaneously. The retention time of proteoforms with large mass shifts can be dramatically different based on the overall properties of each proteoform, including size and hydrophobicity. For quantitative studies, these factors can complicate downstream data analysis; however, they may be overcome by using direct injection of a relatively pure sample (e.g., immunopurified proteoforms) instead of analysis by LC.

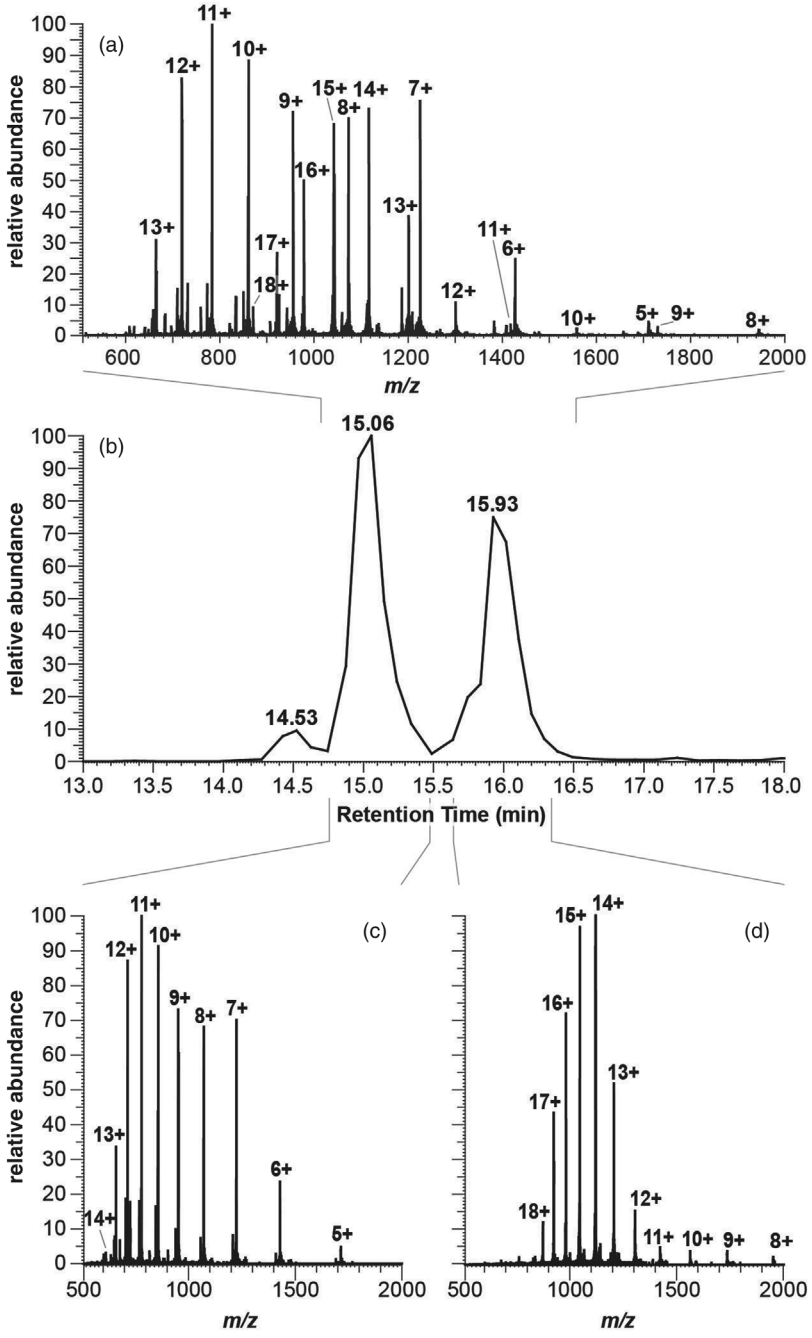


Figure 39.12 High-throughput analysis of complex samples by top-down LC-MS. (a) Overlapping charge states for multiple proteins. (b) Reverse phase liquid chromatography is capable of separating each of the distinct proteins. The charge state distribution is shown for (c) ubiquitin and (c) superoxide dismutase.

Closing Remarks and Perspective In conclusion, top-down proteomics offers a powerful method for the direct analysis of intact proteoforms. Proteins are the central intermediaries between genotype and phenotype. It is not possible to completely understand a biological system if one does not know what protein molecules are present and in what amounts. Given a complex mixture of proteins in a biologically relevant sample, one should be able to determine their precise molecular identities and amounts. Mass spectrometry is currently the best method for this undertaking, also known as "discovery proteomics." Eventually it will also be critical to know where and when proteins are expressed in living systems and the partners with which they interact. Current proteomic approaches provide neither adequate nor complete information on the identities and amounts of whole proteins.

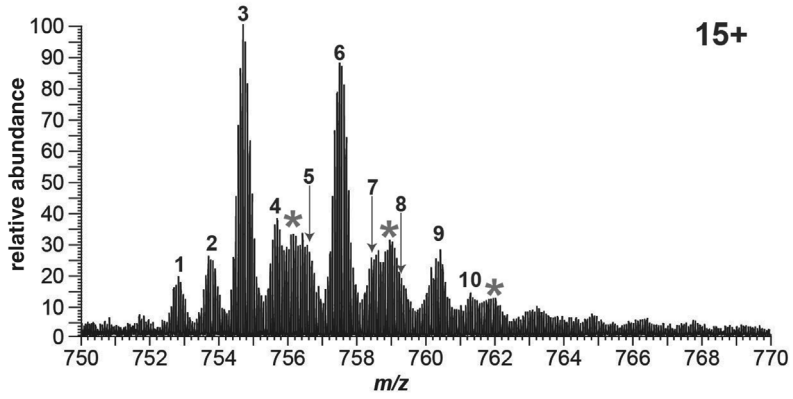


Figure 39.13 Histone H4 proteoforms identified in the 15+ charge state. Ten Histone H4 proteoforms were detected and characterized by isolating each ionic species, fragmenting, and detecting product ions. (See Pesavento, J.J. *et al.* (2008). Combinatorial modification of human histone H4 quantitated by two-dimensional liquid chromatography coupled with top-down mass spectrometry. *J. Biol. Chem.*, **283** (22), 14927–14937). The asterisks represent Histone H4 proteoforms with salt adducts.

The dominant bottom-up paradigm sacrifices information on proteoforms by cleaving proteins into small fragments. This is done for valid pragmatic reasons – it works at scale and leverages a very large investment by academicians, government agencies, and industry. At present, our ability to identify and quantify intact proteoforms is in an early maturation phase. As in genomics, a large effort focused on protein primary structure would drive down the cost per proteoform, just as the cost per base has dropped many orders of magnitude in the past 15 years. The "€1 per base" goal in the mid-to-late 1990s was a powerful impetus, mobilizing the private and public sectors. We envision a similar evolution in proteomics. Although not presently visible, a plan may emerge in the years ahead that embraces the principles of molecular analysis described above for proteoforms and whole protein complexes. This will make complex proteomes far more defined and interpretable – and allow the development of complementary technologies to target the now-known molecules for non-MS-based measurements. The overall effect would be to make the development of diagnostics and therapeutics far more deterministic and efficient. Evidence to this effect is already visible (for additional insights see Kelleher, N.L. *et al.* (2014). Deep and quantitative top-down proteomics in clinical and translational research. *Expert Rev Proteomics,* **11** (6), 649–651).

To drive biological and medical research, we imagine the expansion and development of powerful new technologies that can rapidly identify and quantify the proteoforms present in complex mixtures. These tools may involve mass spectrometry (top-down proteomics), but may also use affinity ligand technologies, emerging technologies, or other new approaches. To realize these goals using mass spectrometry, we advocate the use of intact proteins (and, eventually, protein complexes) as the primary target for measurement in MS-based proteomics and seek to increase throughput via this approach by 1000-fold over current levels. The achievement of the 1000× increase in throughput will require improvements in our current platforms, including separation technologies, ionization methods, mass analyzer ion capacities, multiplexed detector geometries, advanced signal processing, and bioinformatics. In addition, one could readily establish a database containing the identities of all detectable proteoforms that exist in healthy humans. This database will include genetic variants, splice variants, codon substitutions, RNA editing changes, post-translational modifications, and other sources of protein variation in a limited sampling of the human population. We estimate that such a database may contain as many as one billion or more entries (Kelleher, N.L. (2012) A cell based approach to the human proteome project. *J. Am. Soc. Mass Spectrom.,* **23** (10), 1617–1624). This database should be an open data repository to support data collection, analysis, and interpretation. Notably, the Consortium for Top Down Proteomics (http://www.topdownproteomics.org) has begun to assemble such a proteoform repository.

We note that in the field of genomics a difficult discovery phase (imagine the initial sequence analysis of the human genome) was followed by an explosive phase of deep- and re-sequencing presently flourishing in the community. Similar progress from discovery to targeted approaches will provide analogous advances for the development of powerful, widespread, and effective MS and non-MS tools for proteoform analysis. The proteoform database is thus analogous to the genome sequence.

39.5 Peptide Based Quantitative Proteome Analysis (Bottom-Up Proteomics)

H. Alexander Ebhardt, Hannes Röst, Betty Friedrich, and Ruedi Aebersold

ETH Zürich, Institute of Molecular System Biology, Auguste-Picard-Hof 1, 8093 Zürich, Switzerland

39.5.1 Introduction

To describe a cellular system at the molecular level different types of analytes have been identified and quantified. Among these, proteins are particularly informative because they catalyze and control most biochemical reactions of the cell. Proteins are translated from mRNA, are frequently modified post-translationally, and organize in protein complexes to modulate their functionality. Further, the cellular proteome also changes, in response to endogenous and/or exogenous cues, for example, by adapting the protein levels, PTM (post-translational modification) landscape, and organization of proteins into macromolecular complexes. Endogenous cues are, for example, developmental programs or cell cycle progression, while exogenous cues include viral infections or physical or chemical stresses or stimuli. To quantify changes in protein abundance, composition of protein complexes, and/or post-translational modifications, high mass-accuracy mass spectrometry has been used as one of the main techniques in the field of proteomics. However, currently there is no single proteomic method capable of identifying and/or quantifying the entire complexity of the cellular proteome in one experiment. To address some of the analytical challenges posed by the proteome, multiple mass spectrometric methods and strategies have been developed. This section will focus on bottom-up proteomics, an approach in which proteins are isolated from a biological source and digested into peptides using sequence specific proteases. The resulting peptide mixture is then fractionated using liquid chromatography, peptides eluting from the column are ionized by electrospray ionization (ESI), and the resulting peptide ions (also known as precursor ions) enter the mass spectrometer, where different ion species are separated in the gas phase, fragmented in a collision cell, and product ion spectra are recorded (ESI-MS, Figure 39.14).

Bottom-up proteomics identifies and quantifies peptides. However, higher order information such as protein complex composition or sub-cellular localization is generally lost. In addition, the method generates data that reflect the average measurement of a number, minimally thousands but more commonly millions, of cells because single cell measurements are presently beyond the reach of the method. Despite these limitations, bottom-up proteomics is the most frequently applied proteomics method and in combination with other analytical methods (e.g., antibody enrichment of proteins or enrichment of post-translational modifications) it provides a wealth of information to describe a complex cellular system.

39.5.2 Bottom-Up Proteomics

The process of bottom-up proteomics, generically described above, identifies and quantifies peptides that result from the proteolytic digest of proteins. Numerous variations of the general method have been described in the literature and they are not systematically reviewed here. Most implementations share the following features:

1. *Proteolysis using trypsin*: The most frequently used protease is trypsin, which cleaves C-terminal to Arg or Lys residues, and the resulting peptides will therefore terminate with a C-terminal Arg or Lys. The frequency at which these basic amino acids occur within an average protein results in numerous peptides per protein, and these peptides generally have properties with respect to ionization, solubility, and size that are favorable for their analysis by ESI-MS.
2. *Peptide separation*: After protease digestion and purification, peptides are separated based on their relative hydrophobicity using nano-flow liquid chromatography (LC). For ESI-MS, the LC system is directly coupled to a mass spectrometer and eluting peptides are exposed to a differential potential.

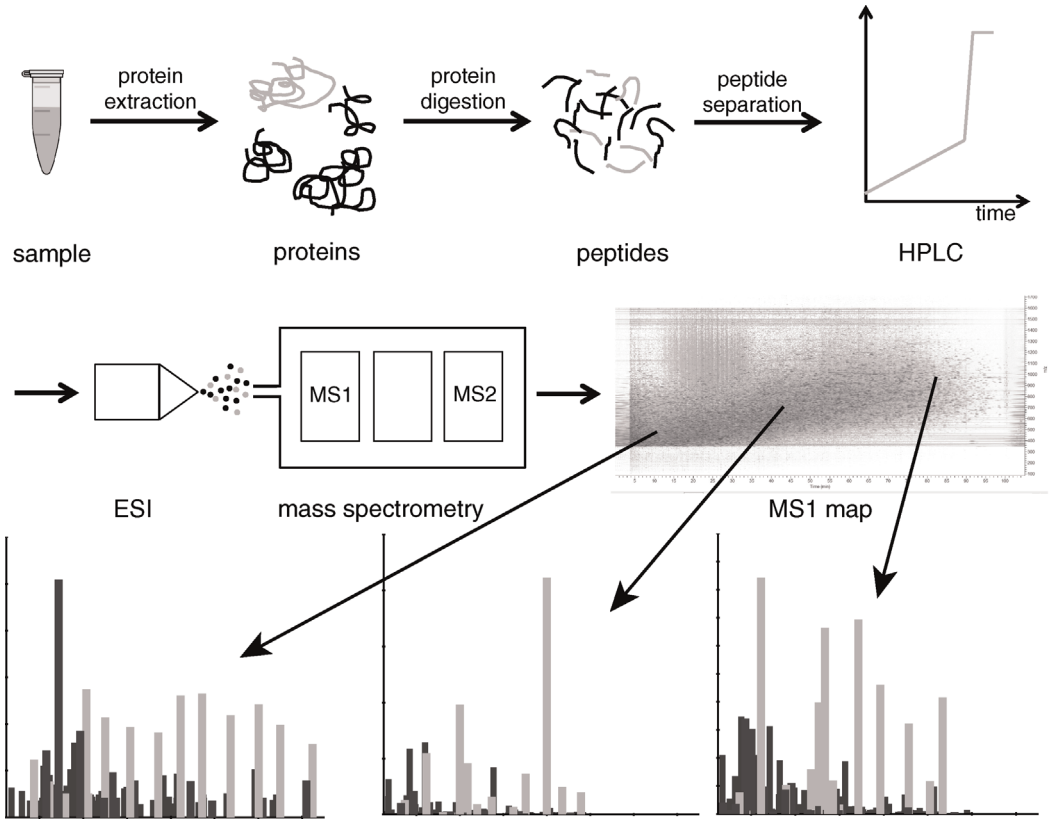


Figure 39.14 Bottom-up workflow for LC-MS/MS. From the initial sample, proteins are isolated and digested using sequence-specific proteases. The purified peptides are typically separated by reverse phase liquid chromatography (HPLC) that is directly coupled to a mass spectrometer. The peptides eluting off the HPLC are ionized and enter the mass spectrometer. Subsequently, the precursor's *m/z* is determined, then fragmented and the MS/MS spectra recorded as a function of *m/z* and intensity. Using *in silico* search engines, the MS/MS spectra are annotated according to a peptide database. As seen in the bottom three MS/MS spectra, the dominant fragment ions are annotated as fragment ions of the precursor.

3. *Ionization*: The ionization potential (in the order of 3000 V) applied at the outlet side of the chromatography column mediates the ionization process of the peptide during which the peptide secures one or more proton charges on basic sites of the amino acid side chains, or the amino terminus of the peptide backbone, and enters the gas phase. At this point the peptide has been converted into a molecular ion, also referred to as precursor ion. The thus generated precursor ions enter the mass spectrometer, which is kept at high vacuum to minimize ion–ion collisions and avoid gas-phase chemical reactions.
4. *Mass measurement and fragmentation*: The mass spectrometer determines, by use of a range of physical devices that guide, filter, store, and fragment selected precursor ions, the mass over charge ratio (*m/z*) of the precursor, and the (*m/z*) value of product ions. Fragmentation is achieved by one of several fragmentation modes, including collision induced dissociation (CID), electron transfer dissociation (ETD), or higher-energy collisional dissociation (HCD).
5. *Data analysis and interpretation:* The product ions spectrum of a particular precursor is a unique fingerprint of the respective peptide and can be manually or computationally analyzed to explicitly read or infer the peptide amino acid sequence. Overall, the whole process, encompassing peptide separation, ionization, precursor fragmentation, and the recording of fragment ion masses and intensities is termed liquid chromatography tandem mass spectrometry (LC-MS/MS). The net result of an LC-MS/MS measurement can be considered as a two-dimensional plot of precursors where the *x*-axis indicates the chromatographic retention time, the *y*-axis the *m/z* value, and the features of the plot are precursor ions detected by the mass spectrometer (Figure 39.14) and fragment ion spectra of precursors selected for fragmentation according to specific rules discussed further below.

All peptide-based proteome analysis strategies start from peptides. In homologous protein species (isoforms, degradation products, or proteins with post-translational modifications), in large numbers identical peptides emerge in the cleavage. Since these peptides are derived from different protein species, they no longer provide the quantitative relations reflected on the protein level. The measured amount of a peptide shows only the sum of all protein species present in the proteome that contains this peptide.

Protein species that show great sequence homology or carry post-translational modifications are usually easy to distinguish in 1D or 2D gels, but are difficult to, or impossible to, distinguish with peptide-based approaches as most peptides are identical.

39.5.3 Complexity of the Proteome

Using current mass spectrometers and extensively fractionated cell lysates, the current number of proteins detectable by LC-MS/MS in a human cell is about 13 000 gene loci. Alternative splicing of mRNA, post-translational modifications (PTMs) of amino acids, and additional mechanisms including protein processing further increase the complexity of the proteome by at least an order of magnitude. The proteolytic digest of these proteoforms using sequence-specific proteases results in a high number of ionized features – estimated to be in the millions – in a typical LC-MS/MS experiment. Because this sample complexity exceeds the capacity of mass spectrometers to sequentially select all precursors for fragmentation, multiple approaches have been described to reduce the complexity of a proteome sample. Proteome samples have been fractionated using prior biological knowledge, for example, the presence of proteins in specific subcellular structures or well-defined protein complexes that can be selectively isolated prior to mass spectrometric analysis. Additionally, proteome samples have been fractionated by using specific post-translational modifications of amino acids as a chemical handle for enrichment. This has been particularly successful in the case of N-glycosylation of asparagine (Asn), phosphorylation of serine (Ser), threonine (Thr), or tyrosine (Tyr) residues, or acetylation of lysine residues. As an alternative to chemical/biochemical fractionation of complex proteome samples upstream of their mass spectrometric analysis, several gas-phase ion gating and fractionation methods have been described that take place within the mass spectrometer and result in populations of fragment ion spectra that have unique properties for each one of these methods. The different bottom-up strategies and their performance are described in the following.

39.5.4 Bottom-Up Proteomic Strategies

There are three main bottom-up approaches to analyzing the proteome: data dependent acquisition (DDA), targeted methods such as selected / multiple / parallel reaction monitoring (SRM, MRM, PRM), and data independent acquisition (DIA).

Mass Spectrometry, Chapter 15

Principle of DDA In DDA the complete range of *m/z* covered by the mass spectrometer is surveyed as a function of time, resulting in MS1 values for all co-eluting precursors. This so-called "survey scan" generates a list of the most intense ions from which the 10–20 most intense precursors are selected for subsequent MS/MS fragmentation. Once complete, another survey scan takes place and the cycle repeats. This iterative process takes between 1 and 2 s, depending on exact specifications and covers a wide range of precursors that will later be annotated as peptides. DDA is an unbiased approach to sampling the peptide sample, but contains a stochastic element for selecting the precursor to be fragmented (Figure 39.15b). Specifically, if the number of precursors detected in the survey scans substantially exceeds the number of sequencing cycles that the mass spectrometer can execute in the time available, in repeat analyses of the same sample different precursor populations are selected and therefore different sets of peptides identified. DDA is most effectively used as a discovery technique.

Principle of SRM Selected reaction monitoring, in contrast, is a biased approach to sampling the peptides in a sample. It consistently monitors the same number of approximately 100 peptides

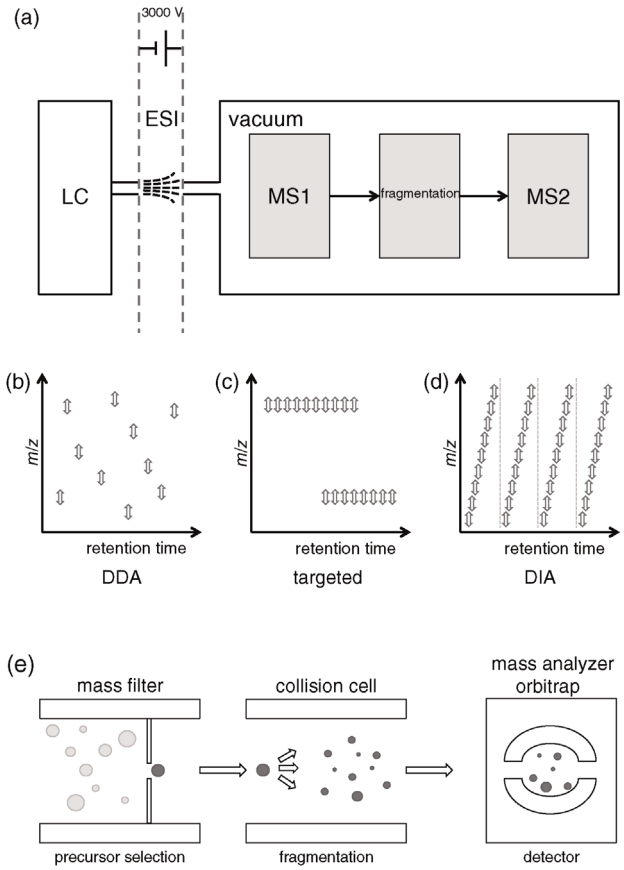


Figure 39.15 Mass spectrometer basic principles. (a) Basic components of a mass spectrometer: the high pressure reverse phase liquid chromatography (HPLC) column is directly coupled to the mass spectrometer. Eluting peptides are exposed to a differential potential (approximately 3 kV). Within this electrical field peptides turn into ions and are called precursors, hence the term electrospray ionization (ESI). The precursors enter the mass spectrometer, which is kept under vacuum, to determine the precursor's *m/z*. Subsequently, the precursor is fragmented and all precursor specific product ions are measured again. (b)–(d) Diagrams plotting selected precursors (*m/z*) as a function of time for (b) data dependent acquisition (DDA), (c) selected reaction monitoring (targeted), and (d) data independent acquisition (DIA). Note, when comparing multiple LC-MS/MS runs, DDA might not always identify/quantify the same peptide due to the stochastic sampling approach. This is in contrast to Targeted, where certain *m/z* values are consistently measured, but many peptides remain unquantified. DIA consistently covers the entire *m/z* range as a function of time. (e) Typical DDA mass spectrometer. First, specific precursors are selected for fragmentation. Upon fragmentation, the fragment ions are measured in a mass analyzer, (e.g., Orbitrap (shown) or by time of flight).

per LC-MS/MS experiment across many samples by specifically targeting those peptides for analysis in each sample. The set of targeted peptides needs to be specified prior to each analysis and for each targeted peptide specific coordinates such as the retention time and the fragment ion spectrum need to be predetermined. SRM is the most sensitive LC-MS/MS method currently commercially available and can be standardized between laboratories easily with very small coefficients of variance. SRM overcomes the stochastic sampling issue in DDA, but is limited by the number of peptides quantifiable per sample (Figure 39.15c).

Principle of DIA Independent of any prior information, DIA fragments precursors and records the *m/z* values of all fragment ions. An implementation of this approach is SWATH-MS (sequential window acquisition of all theoretical masses), in which sequential swathes of 25 *m/z* are fragmented and all fragment ions *m/z* values determined using a time of flight (ToF) mass analyzer. Current instrumentation scans the *m/z* range from 400 to 1200, which is covered in 32 swathes. Each cycle of fragmenting the precursors in a set of sequential swathes takes about 3.2 s and is iteratively repeated until the LC gradient completed (Figure 39.15d). The net result of SWATH-MS and similar DIA approaches is a complete map of fragment ions for each precursor generated from a sample. Because, generally, multiple precursors are present and concurrently fragmented in each extended precursor ion selection window used for data acquisition, the fragment ion spectra are convoluted and the identification of peptide sequences from such spectra has been challenging. Below we describe reliable computational strategies for the identification and quantification of peptides from SWATH-MS datasets.

39.5.5 Peptide Quantification

Simply identifying proteins within a complex sample will rarely lead to new insights into the biological system as biological systems react to exogenous or endogenous cues with

quantitative changes. Therefore, quantification of the identified peptides in bottom-up proteomics can provide useful information about the dynamic of cellular responses. There are two general methods to quantify peptides: label-free or based on stable isotope labeled reference peptides. Label-free quantification uses the integrated area under the curve of a specific feature in the *m/z* versus retention time map and is frequently used to quantify relative abundance changes in similar samples, exemplified by the time course of perturbation experiment. Stable isotope labeled reference peptides take advantage of naturally occurring ^{13}C and ^{15}N stable isotopes, are chemically synthesized, and can be externally calibrated using amino acid analysis. These heavy peptides, which have retention times identical to those of their light counterparts, are used to positively identify peptides and serve as reference standards for quantification. Besides chemical synthesis of ^{13}C and ^{15}N peptides, heavy isotopes can be incorporated into proteins by metabolic labeling through chemically modifying proteins after synthesis (Section 39.4.3). Metabolic labeling is frequently achieved by substituting essential amino acids in the growth medium of cells with stable isotope labeled variants (e.g., Lys $^{13}C_6$, $^{15}N_2$ or Arg $^{13}C_6$, $^{15}N_4$). This approach is also technically feasible on an organismic level (e.g., mouse). For protein labeling with heavy stable isotopes by metabolic labeling of by chemical reactions numerous protocols have been developed which are not further reviewed here (Section 39.4.3).

Peptide quantification using DDA can be performed using label-based and label-free methods; however, as feature quantification is based on MS^1 level signals, there are frequently interferences that may mislead quantification. Unique to DDA is the count of annotated MS/MS spectra, called spectral counting. As the name suggests, all spectra positively identifying a peptide are counted towards the protein total count. The resulting number is integrated over the length of a protein and used as proxy for protein abundance.

Quantification by SRM is based on MS^2 level signals and therefore less prone to interference compared to MS^1 level quantification. The most accurate quantification is achieved by SRM when stable isotope labeled reference peptides are used, but label-free quantification has also been demonstrated, particularly in cases in which a large number of peptides was quantified per LC-MS/MS run.

Similar to SRM, quantification in peptides in SWATH-MS is usually on the MS^2 level. As the number of identified and quantified features is in excess of 10 000 per LC-MS/MS run, most SWATH-MS experiments are quantified using label-free approaches where signal intensities are normalized to the total amount of peptides or ions detected.

39.5.6 Data Dependent Analysis (DDA)

DDA Principle and Intended Use Data dependent analysis requires a survey scan of all precursors co-eluting at a specific time, followed by the sequential selection of detected precursors for subsequent fragmentation and the recording of the resulting fragment ion spectra (Figure 39.15e). Most frequently DDA experiments are carried out on orbitrap mass spectrometers, although other forms of tandem mass spectrometers are also well suited to perform the method. All MS^1 and MS^2 information is assigned to their corresponding peptide sequences by means of *in silico* search engines. The ensemble of peptides identified in a measurement is then used to infer the population of proteins that is best represented by the peptides. DDA is often referred to as discovery proteomics as its main goal is the identification of proteins present in a particular sample. The dynamic range of proteins/peptides covered by DDA is three to four orders of magnitude.

DDA Strength and Weaknesses DDA is a well-established method and its maturity both in terms of the mass spectrometry instrumentation and *in silico* search engines for MS^2 spectra annotation is a strength of the method. The weakness of DDA derives from its stochastic sampling of precursor ions, which leads to poorly reproducible sets of identified peptides if complex samples are repeatedly being analyzed. These effects of undersampling can be reduced if complex samples are extensively fractionated and the generated fractions are sequentially analyzed. However, multiple LC-MS/MS runs required for the analysis of a fractionated sample are quite time consuming.

Typical Applications of DDA Using DDA, many cellular systems have been characterized extensively, ranging from microbial proteomes over differences between haploid and diploid yeast genomes to the survey of several human cell lines and other human specimen. To reach deep coverage of these proteomes extensive fractionation of the isolated proteomes is usually applied. Discovery proteomics has built a huge inventory of peptides identifiable in these samples. DDA combined with enrichment methods, such as affinity purification mass spectrometry (AP-MS) has led to important discoveries of protein complex compositions, which frequently change as a function of cues. Besides protein identifications of various samples, DDA has been instrumental in identifying PTMs of proteins such as, for instance, phosphorylation or poly-ubiquitination.

DDA Challenge Of critical importance for the understanding of biological systems is the knowledge of how the identified peptides behave within the cellular system as a function of endogenous and/or exogenous cues. For example, how does the proteome of a human cell change as a function of viral infection? To answer such questions, typically a time or dosage course covering a viral infection of a human cell starting with uninfected cells and several time points past infection and/or multiplicities of infection is set up. If ten time points were to be measured with biological replica at three different multiplicities of infection at triplicate repeats one easily arrives at 100 samples for bottom-up proteomics in a single study. Extensive analytical fractionation to maximize peptide identification is in this case not practical, as the number of LC-MS/MS runs would increase linearly with the number of fractions. Further, the stochastic sampling component of DDA LC-MS/MS leads to "holes" in the data matrix, making interpretation of biological data extremely challenging if a large number of samples has to be quantitatively or qualitatively compared. Therefore, if highly reproducible proteomic datasets are required in a study, targeted LC-MS/MS approaches are frequently applied.

39.5.7 Selected Reaction Monitoring

SRM Principle and Intended Use Selected or multiple reaction monitoring (SRM or MRM) is a targeted bottom-up proteomics method. The main objective of the method is the accurate quantification of a set of pre-selected proteins in a sample cohort. SRM uses triple quadrupole mass spectrometers, which are fast, robust, very sensitive, and the dynamic range spans at least five orders of magnitude (Figure 39.16a). Various cross laboratory studies showed that standardized measurements can be carried out across many types of SRM mass spectrometers with achievable CVs (coefficients of variation) of 5% making SRM the current "Gold Standard" method for quantitative proteomics.

Principle of SRM Targeted proteomics is based on prior knowledge for each targeted peptide. Specifically, the peptide's relative hydrophobicity that determines its retention time, the precursor ion mass, and the fragmentation pattern of the precursor ion collectively constitute a mass spectrometric assay that confidently determines the presence of the targeted peptide in a sample. The prior knowledge of peptides is typically obtained using DDA LC-MS/MS. The parameters of the assay are vital for targeted proteomics, as SRM is based on first filtering the correct m/z of the precursor using electrostatic fields of the first quadrupole, followed by the fragmentation of the selected precursor(s) in the second quadrupole, and the selection of signals for specific fragment ions by the third quadrupole. Ions passing these sequential filters are recorded by a detector, but no actual m/z values are measured. Hence, the specificity of SRM relies on the sequential filtering capabilities of the quadrupoles (Figure 39.16a).

A transition is defined as the pair of m/z values for precursor and fragment ion. Per peptide, multiple transitions are measured at consecutive retention times, which are typically spaced in time such that 8–10 measurements are recorded for each transition of a targeted peptide across the chromatographic elution time. The resulting signals are used to reconstruct chromatographic elution profiles for each transition. The resulting peak groups are the basis for the unequivocal identification and accurate quantification of a targeted peptide. For quantification, the area under the curve per peak group is integrated and used as an indicator of the peptide's quantity. By scheduling SRM methods (defining a time interval in which each SRM assay is measured), about 100 peptides can be quantified per SRM LC-MS/MS run. Figure 39.16b illustrates a

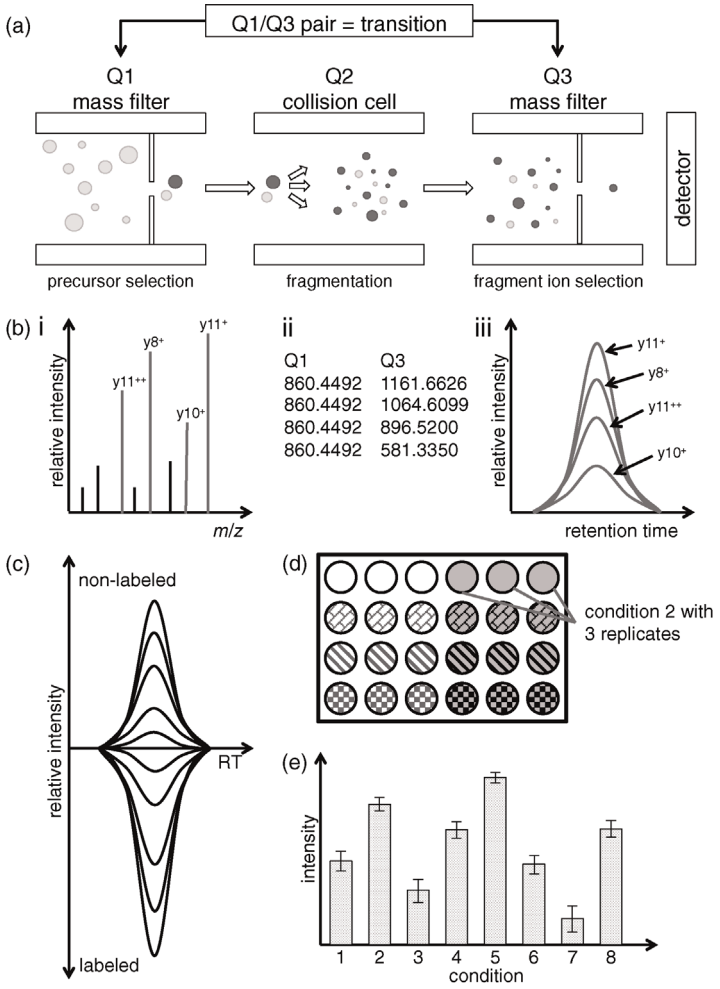


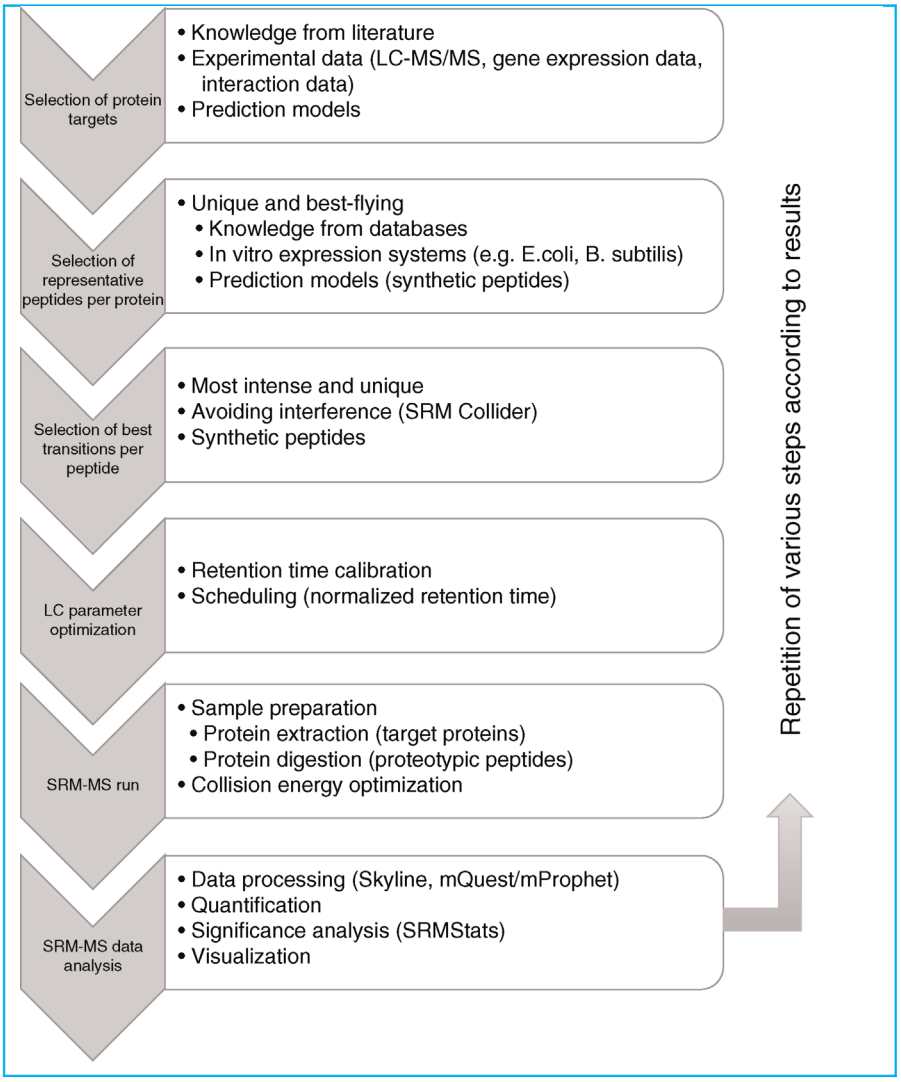
Figure 39.16 Targeted proteomics. (a) Principle components of a triple quadrupole (Q) mass spectrometer. Precursor ions matching a preset *m/z* value (±0.35) pass the first Q, and are fragmented in the second Q. Fragment ions matching a preset *m/z* value (±0.35) pass the third Q and are measured by the detector. As this detector does not measure *m/z* values, but only ion current, values measured depend on the specificity of the previous two Q filter. (b) Prior knowledge is required to quantify peptides in Targeted proteomics. (b, i) From a typical MS/MS spectrum obtained from DDA experiment, the most intense fragment ions are selected. (b, ii) Per precursor, typically four product ions are measured. Each Q1/Q3 pair is called a transition and all transitions to be measured are entered into a method table for the QQQ mass spectrometer. (b, iii) All transition within the method are measured as a function of time resulting in an ion chromatogram. (c) For accurate quantification of peptides, especially in large datasets, stable isotope labeled reference peptides are spiked into the sample and quantified across all samples. As these reference peptides share identical LC and ESI properties with the non-labeled peptide (but not *m/z*) both peptides co-elute at the same retention time. (d) Mass spectrometers used for targeted proteomics have a very low CV of less than 20% as shown in many cross-laboratory studies, with most variability originating from the biological system and sample workup. Hence, the use of biological triplicate measurements is currently standard in Targeted proteomics for subsequent statistical analysis of the data as seen in (e).

MS/MS spectrum obtained using DDA LC-MS/MS (left-hand panel), a list of transitions to be measured in the middle, and the resulting ion chromatogram in the right-hand panel.

Method The flowchart shown on the next page in the box describes the SRM method development, an iterative process that starts with a list of proteins. This list of candidate proteins is derived from prior knowledge or *in silico* prediction algorithms. As SRM is a bottom-up proteomics approach, appropriate peptides corresponding to each candidate protein are selected from databases collecting SRM assays such as PASSEL (Peptide Atlas SRM Experimental Library) or Panorama that contain large numbers of DDA spectra. If no prior empirical knowledge exists, SRM assays can be generated rapidly using bacterial or eukaryotic protein overexpression systems. The overexpressed proteins are purified, digested with a protease, and generated peptides identified by DDA ESI-MS. This workflow will result in empirical evidence of best performing peptides. Notably, the digested overexpressed protein approach will result in equimolar peptide ratios, but there will be only three to five peptides per protein with very favorable ESI-MS properties when compared to other peptides of the same protein. Hence, the alternative of using an *in silico* algorithm that predicts proteotypic peptides with favorable ESI-MS properties should be used as last resort.

Following the peptide selection process the next step is the selection of most intense and unique transitions. Similar to the best performing peptides per protein, there are best performing transitions per peptide, which are specific to a certain type of SRM mass spectrometer. In addition, due diligence should be taken to avoid interfering ions not originating from the peptide

of interest. This is especially critical for very complex samples such as undepleted human cell lysates. There are *in silico* prediction tools that will calculate interferences (e.g., SRM Collider). Interferences are also apparent in the actual SRM ion chromatograms. Chemically synthesized peptides are also a viable alternative to optimizing SRM measurements. Another important aspect of SRM method optimization is the retention time calibration for subsequent scheduling. The relative retention time per peptide is determined relative to other endogenous peptides or relative to spiked in synthetic retention time peptides. Last, but not least, the collision energy (CE), which is the energy value used to fragment a precursor in the collision cell, needs to be optimized. Using a step-wise CE ramp, the transitions are measured as a function of CE to maximize the signal of each transition. Once optimized, a scheduled SRM method reliably quantifies about 100 peptides per injected sample.



Identification Only in very few cases is a single transition sufficient to uniquely identify a peptide analyte in a complex sample. However, if multiple transition signals that precisely co-elute at a specific retention time are used to identify an analyte, it becomes increasingly unlikely that another peptide shares all of the target peptides transitions and could thus produce a similar-looking peak group during data acquisition. In general, researchers use a mixture of theoretical and heuristic approaches to ensure the correct identification of the target signal. These approaches can help to distinguish between multiple peak groups that are visible in the data and help to resolve

any ambiguity regarding which peak group is the signal correctly identifying the targeted peptide. In practice, researchers use multiple, orthogonal pieces of information to decide which signal represents the true analyte:

1. *Signal quality:* If the individual transition signal traces were truly generated by a single peptide analyte eluting from the chromatographic column, the traces should be identical in shape and co-elute exactly. The researcher can use this information to assess peak quality and remove peak groups whose individual traces are dissimilar in shape (i.e., if some peaks are broader than others) or do not exactly co-elute (i.e., if there is a distinct shift in the peak apex between traces). Owing to the nature of the method, where the analyte only gets fragmented inside the mass-spectrometer, peaks not exhibiting these quality signs can often be discarded.
2. *Signal correspondence to a priori knowledge:* The fragmentation method and the chromatographic separation used in LC-MS/MS are highly reproducible processes and, therefore, the elution time and the relative intensity of the individual fragment ions (traces) should be very similar between different runs. Often the analyte assay is generated from empirically recorded peptide fragmentation events and the normalized intensity of the fragment ions is therefore known experimentally (often the intensity is normalized relative to the most intense ion). If the intensities of the traces observed in a peak group are very different from the expected ones (i.e., the relative order is not preserved), this is indicative of a noise signal. Similarly, the chromatographic retention time is often known along with the fragment ion intensities and if normalized properly (e.g., using a set of standard peptides spiked into every run) can be used to identify the correct peak group.
3. *Signal correspondence to a spike-in standard:* Stable-isotope-labeled peptide standards that are identical to the target peptide are often used in SRM experiments for exact quantification but can also serve during the identification phase. Since the stable-isotope-labeled peptides are identical to the target peptide in all their physiochemical properties except their mass, they will also show identical behavior in the chromatography (Figure 39.16c). If a non-labeled group and a stable-isotope-labeled peak group co-elute at the exact time point and have the same relative fragment ion intensities, it is very likely that this signal is produced by the true analyte. Especially with complex backgrounds or large-scale clinical studies, using a spike-in isotopically labeled peptide standard helps to confirm the true identity of the measured analyte.

Quantification Once the analyte has been identified, the next goal in quantitative proteomic experiments is to obtain either absolute or relative quantification across several biological samples (Figure 39.16d, e). For most use cases, relative quantification is sufficient. For example, to determine differentially abundant proteins between healthy and diseased individuals in a clinical study, or to determine the reaction of a biological system to an external perturbation (such as environmental changes, pathway activation, drug delivery, etc.) relative quantification between samples is frequently employed. Relative quantification can be obtained by directly comparing the absolute signal intensity of the different samples. Alternatively, absolute analyte quantities can be determined via an internal standard of isotopically labeled reference peptides. For certain mathematical models in systems biology, absolute protein quantities are required and in addition they allow further insight into the energy expenditure and the relative importance of different protein classes for the organism. In the following we discuss different quantification strategies:

1. *Label-free quantification:* This strategy does not depend on spike-in peptides but directly uses the ion counts from the mass spectrometer to compare the relative abundances of analyte counts across different samples. Several experimental factors such as the sample amount, time of analysis, instrument condition, and sample background may influence the reported ion counts and may introduce biases in the quantitative analysis. The label-free strategy is viable if the background is sufficiently similar across samples tested and the sample preparation is reproducible, which needs to be ensured through independent replicate experiments.
2. *Stable isotope-based quantification:* In this strategy, for each targeted peptide a matching stable-isotope-labeled peptide (or protein) is added to each sample in equal amounts. This

allows us to control and remove many of the aforementioned biases and sources of variation since any change in reported ion counts will affect the endogenous and the labeled peptide equally. It is therefore important to add the internal standard as early as possible in the sample preparation workflow to reduce variation. Isotopically labelled proteins can be added before tryptic digestion and can also account for incomplete or unspecific proteolysis. In contrast, labeled peptides are typically added after proteolytic digestions. During analysis, each peptide is quantified relative to its matching stable-isotope labeled peptide and these relative values are then compared across samples. Assuming the same amount of stable-isotope peptide was added to each sample, these values then allow direct conclusions to be drawn about abundance changes in the studied samples.

3. *Absolute quantification:* If the absolute concentration of the stable-isotope-labeled spike-in peptides or proteins is known, the ratio between the ion counts of the targeted endogenous peptide and the stable-isotope-labeled peptide allows us to directly compute the concentration of the targeted peptide in the initial sample. This can be achieved if the spike-in peptides or proteins are obtained in analytically pure form and have their concentration determined by analytical techniques such as quantitative amino acid analysis. To determine the biological concentration of a protein in a sample, the amount of sample (e.g., number of cells or volume of a body fluid) used to generate the sample is related to the quantity determined for the analyte in question to calculate $\mathrm{mg\,ml^{-1}}$ or copies per cell concentration values. The accuracy of the technique depends on the accuracy of the concentration of the spike-in peptides, as well as the sample loss before the spike-in peptides were added to the sample. To increase the accuracy of quantification, the peptides or proteins should be added as early as possible in the process.

SRM Data Analysis In targeted proteomics data analysis is an important step in the workflow and needs to be properly accounted for. The data produced in targeted proteomics is visualized in Figure 39.17a and is generally acquired in multiple *chromatographic traces,* each one representing a transition signal acquired over chromatographic time that form distinct *peak groups* at positions of potential peptide analyte elution. The task during data analysis is to group all traces belonging to the same analyte together and to determine whether the resulting peak group correctly identifies the targeted analyte, thus indicating the presence of the peptide analyte in the sample.

Even though in SRM two mass filters are applied to achieve high specificity, a particular transition signal may not be specific for a particular peptide, especially in a complex sample (Figure 39.17b). Other peptides with similar precursor and fragment masses may produce such non-specific signals. These peptides might either produce identical precursor and fragment ion pairs due to (partial) sequence similarity or produce similar precursor and product ion pairs, simply by chance, that are indistinguishable by the limited resolution of the quadrupole mass filters. In addition, non-canonical protein isoforms, post-translational modifications, and the natural isotope distribution increase the likelihood of such an event occurring. If no further validation is performed, such signals can easily be mistaken for the true signal. It is therefore important to validate the SRM measurements either manually or by using specialized software (see below).

Analysis Software Analysis software plays an important role in the data analysis of targeted proteomics. The software can support the researcher during all steps of an SRM experiment, starting with the protein and peptide selection to the creation of appropriate measurement schedules for each targeted peptide and, particularly, during data analysis. Some software packages, specifically the popular Skyline system, allow the visual representation of the peptides, proteins, precursors, and transitions and relate those hierarchical relationships to the data acquired on a mass spectrometer. In addition, some software tools, specifically the mProphet software, perform multiple steps of the data analysis automatically, which removes manual bias, improves reproducibility in data analysis, and enables researchers to perform large-scale studies that would not be possible without software. Owing to the many steps involved in a targeted proteomics experiment, often more than one software tool has to be employed.

The proteomics community has produced several open-source software packages that support researchers during a targeted proteomics experiment. Many tools contain a transition

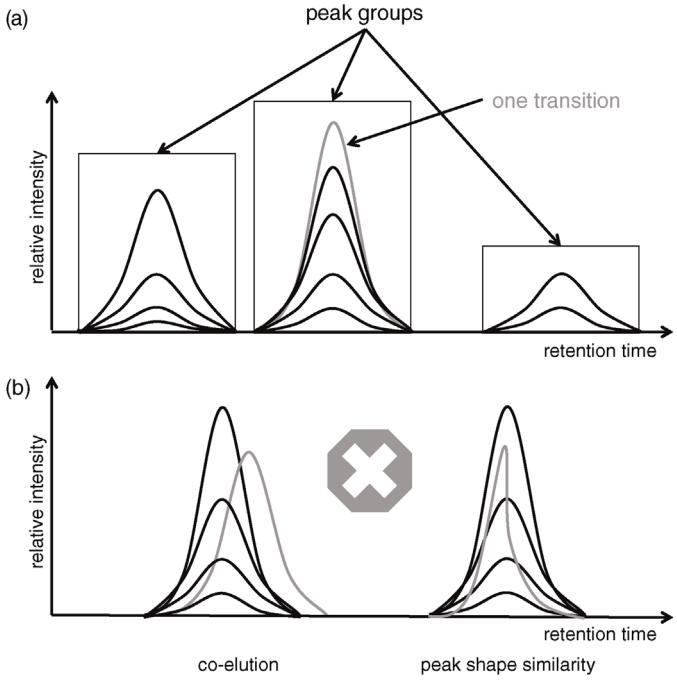


Figure 39.17 Ion chromatograms: (a) Per peptide, at least four transitions are measured that co-elute as peak groups. (b) The detector does not determine the *m/z* of ions passing through the QQQ, but relies on the specificity of the Q1/Q3 mass filters. Hence, it is possible that ions not originating from the intended precursor interfere with the measurements. These interfering ions can be spotted as they do not perfectly co-elute with other transitions of the targeted peptide. Hence, it is important to measure at least four transitions per peptide, which should all precisely co-elute.

selection component that assists in experimental design, a peak-picking module in the chromatographic dimension, and a statistical modeling step that computes a quality score for a specific peptide (e.g., a *p*-value or a *q*-value). In general, the software tools use the same criteria as outlined above for manual validation to evaluate the quality of a peak and often compute a score for each orthogonal quality metric. These individual scores are then combined to produce a single score that indicates the overall quality of a specific peak group. Several statistical and machine-learning techniques have been employed to determine how to combine the individual scores in an optimal fashion such that all high-quality peaks obtain a high score while low-quality peaks are assigned a low score. A statistical analysis of the data then allows the computation of cutoff scores that allow the control of experiment-wide error rates. Specifically, the *mProphet* algorithm first allowed such automated analysis of targeted proteomics data employing linear discriminant analysis (LDA) to differentiate true and false peaks optimally.

One open-source software package that has become increasingly popular is the Skyline software, which allows users of Microsoft Windows to design a targeted proteomics experiment and after acquisition load the resulting data for visualization and peak-picking into the graphical user interface. The easily accessible usability supported by thorough documentation, an active development community, and large number of features make Skyline a popular choice among proteomics researchers for data analysis as it allows the user to easily visualize a multitude of bottom-up proteomics data acquisition methods: from DDA over SRM to DIA (data independent acquisition, covered below).

For optimal assay selection, it is important to create assays that are sequence-unique (proteotypic) as well as transition-unique in the chosen analysis background. This means that there should not be any other peptide in the sample with the same sequence (sequence-unique) or producing the same transition (transition-unique). In addition, it is often helpful to select peptides from a protein that are likely to produce intense signals in the mass spectrometer. To select proteotypic peptides, empirical databases such as the Panorama (panoramaweb.net, MacCoss laboratory), the PeptideAtlas and the related SRMAtlas project allow researchers to access thousands of previous MS runs and select the optimal peptides based on empirical evidence. The PeptideAtlas projects contain predictions regarding visibility by mass spectrometry, hydrophobicity, and links to individual mass spectra if the peptide was previously identified. The SRMAtlas project furthermore contains information about specific assays and allows researchers to interactively explore these assays in the context of actual experiments,

and to download ready-made transition lists. In addition, researchers can also use computational tools to ensure transition-uniqueness of their designed assays. Tools like the SRMCollider (also integrated within Skyline) can search a proposed assay against the background proteome and identify potentially interfering peptides that share multiple precursor/fragment ion pairs, allowing researchers to refine an assay early in the design process and, for example, choose different transitions that are non-ambiguous (unique) in a given sample.

SRM Strength and Weaknesses The consistent quantification of peptides using SRM across a large number of samples is key to confirming a hypothesis with statistical significance. SRM is also the most sensitive LC-MS/MS method currently commercially available. Hence, targeted proteomics using SRM is an advance for quantitative proteomics, but limited to about 100 peptides that can be quantified per LC-MS/MS run.

Typical Applications of SRM Once the targeted bottom-up proteomics method SRM is established, it allows for the most sensitive proteomics measurements to be performed and a consistent set of peptides to be quantified as a function of perturbation, to advance knowledge about the biological system at hand.

Microbial Targeted proteomics can be applied on microorganisms for various purposes. Some of these include elucidation of bacterial mechanisms of virulence and antibiotic resistance, investigation of potential drug targets, and host–pathogen interaction. Specifically in microorganisms that are well-studied and where quantification (as opposed to discovery) is the main interest, using targeted proteomics can allow researchers to study dynamic changes in the microbial proteome with great precision. In the following, we describe several example studies in closer detail.

Schubert and coworkers created a proteome-wide assay library for the human pathogen *Mycobacterium tuberculosis*, which allowed them to investigate its 4000 proteins using targeted mass spectrometric approaches. Using a set of over 15 000 peptides synthesized for this purpose, they demonstrated that mass spectrometry achieves similar coverage as RNA sequencing in discovery mode. They used targeted proteomics to quantify 45 out of 53 proteins in the dormancy regulon of *M. tuberculosis,* providing novel protein-level insights into the regulation of important regulatory factors during dormancy. Other researchers have used targeted proteomics in *M. tuberculosis* to investigate substrates of Esx-1, a virulence factor translocation machinery and to identify mycobacterial peptide sequences in exosomes isolated from human serum samples obtained from patients with active and latent tuberculosis.

Synthetic biology researchers have demonstrated that targeted proteomics can be used as an efficient method to measure protein expression, which has allowed them to optimize metabolomic flux through engineered pathways in *E. coli*. The researchers assessed the expression of proteins in the mevalonate pathway and tyrosine biosynthesis, demonstrating the importance of determining accurate protein levels in engineered systems and to fine tune metabolic pathways. Using the proteomic data, the researchers were able to substantially boost the production of desired metabolites in genetically modified bacteria.

Taken together, these examples show the broad applicability of targeted proteomics in the study of microorganisms. Specifically, targeted proteomics can be applied in any situation where accurate and reproducible protein quantification is desired across high numbers of samples. Since antibodies are often unavailable for microbial proteins, applying targeted proteomics methods is often a cost-effective way to achieve the desired specificity and recent studies on multiple organisms have demonstrated high sensitivity of targeted proteomics, allowing researchers to monitor almost any protein with little effort.

Eukaryotic Model Organisms Eukaryotic model organisms such as the yeast *Saccharomyces cerevisiae* or mouse (*Mus musculus*) have been characterized extensively phenotypically. Besides phenotypes, the genomes of these model organisms are known and many other molecular measurements are available, such as single nucleotide polymorphisms of genomic DNA, transcriptomics data, metabolomics, and, through extensive fractionation, the detectable proteome has been surveyed. Targeted proteomics builds on this prior knowledge and is typically

applied to quantify a predefined set of peptides, and therefore proteins, as a function of perturbation. If this set of peptides is chosen carefully, it can represent the adaptation of a protein network to adjust to different environmental cues or be surrogates representing various cellular states. Other advantages of model organisms are the well-defined genomic background, which often is resolved to the level of single nucleotide polymorphism per strain. One such example of an extremely well-defined model organism is the BDX mice strains. Many studies have used these mice strains to elucidate various molecular mechanisms, including a study using targeted proteomics to dissect mitochondrial activity linking genetic traits to phenotypic appearance as a function of diet.

Clinical The widest range of applications of targeted proteomics is found in clinical studies where a relatively small number of peptides, and therefore proteins, are quantified across many patient samples. Protein candidates targeted by SRM-MS typically arise from exploratory pilot studies and are quantified across a larger number of patient samples (e.g., to validate these proteins for clinical use). For example, Cima and colleagues compared the N-glycosylated peptides in serum of prostate cancer model mice with PTEN knockout to wild-type mice. As a result of this initial comparison, 39 N-glycosylated peptides were chosen as candidates and quantified across 143 serum samples from prostate cancer patients and age matched control group. The results of the targeted proteomics quantification was evaluated by statistical testing and resulted in four N-glycosylated peptides suitable for diagnosis and grading of prostate cancer.

Recent advances in mass spectrometry instrumentation make it now possible to extend the quantitative robustness of SRM to all peptides detectable by the mass spectrometer. The method is called sequential window acquisition of all theoretical fragment ion spectra (SWATH-MS).

39.5.8 SWATH-MS

SWATH-MS Principle and Intended Use SWATH-MS relies on high mass accuracy measurement of fragment ions using a time-of-flight mass analyzer (Figure 39.18a). SWATH-MS does not explicitly target single precursors but rather fragments whole swathes of the precursor ion range simultaneously and then uses downstream software to computationally create ion traces (Figure 39.18b). Generally, the swathes are chosen to cover most precursors (e.g., swathes of *m/z* 25 width covering the mass range *m/z* 400–1200 in 32 steps), although other configurations of precursor selection windows are also applicable. After acquisition, researchers can decide which peptides and which transitions to extract from the dataset and are not limited to the set of recorded precursor/fragment ion pairs.

SWATH-MS data analysis required the development of new software tools, as traditional *in silico* search algorithms are unable to correctly annotate the mixed MS/MS spectra containing fragment ions of multiple precursor ions. SWATH-MS quantifies the complete detectable peptidome. With this extensive coverage, the biological sample is converted into a digital map that allows for MS/MS based quantification. The peptide quantification is carried out on the MS/MS level and together with the large number of peptides quantified makes it possible to use label-free quantification approaches instead of stable-isotope labeled external reference peptides.

The output of SWATH-MS is fragment ion traces in the form of ion chromatograms as a function of time (Figure 39.18c) and is therefore conceptually similar to the data structure generated by SRM. As each precursor fragments into multiple fragment ions, all fragment ions co-elute in the ion chromatogram as a single peak group. One way to annotate these peak groups is to compare them with previously acquired MS/MS spectra from DDA LC-MS/MS experiments as implemented in the OpenSWATH software, which also includes an mProphet module for optimized high throughput called pyProphet. The comparison includes matching of retention time and relative fragment ion intensities together with other discriminating scores, such as peak group shape and matching of peak groups with decoy databases. There are several repositories containing assays of unmodified and modified peptides for interrogating SWATH-MS maps (e.g., ProteomeXchange or SWATHAtlas).

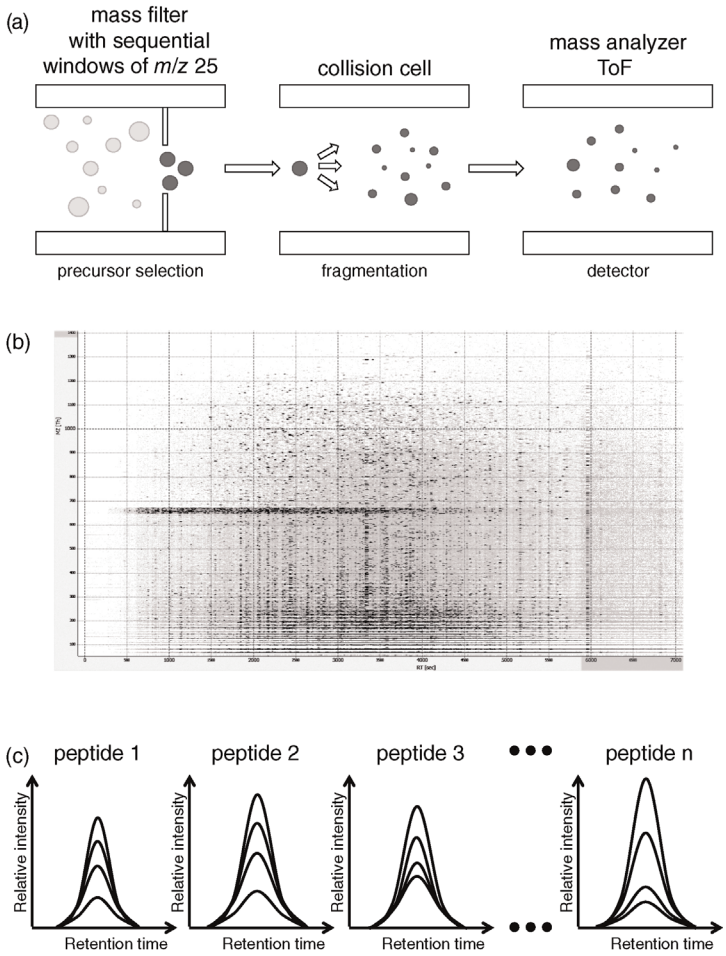


Figure 39.18 SWATH-MS: (a) General principle of the SWATH-MS mass spectrometer. As in SRM, the first mass filter selects precursors; however, in SRM the target *m/z* value has a window of *m/z* ±0.35 while in SWATH-MS this window is *m/z* ±12.5. All ions passing through this first filter are fragmented and the *m/z* values of each fragment ion determined by a time of flight mass analyzer. (b) The sequential acquisition of fragment ions passing through *m/z* ±12.5 precursor windows leads to a complete *m/z* map of fragment ions as a function of time. (c) Using various *in silico* tools, this SWATH-MS *m/z* map is annotated and individual ion traces of co-eluting peak groups visualized, similar to SRM, but per SWATH-MS run more than 10 000 peptides can be quantified in a peptide mix originating from whole cell lysate.

SWATH-MS Strength and Weaknesses The strength of SWATH-MS is its unbiased recording of precursors and precursor fragment *m/z* values with high mass accuracy as a function of LC retention time. This allows for consistent MS^2 quantification of peptides across many samples. A challenge of the method is the potentially increased background compared to corresponding SRM data as a result of the size of the SWATH-MS precursor ion selection window and the size of the resulting files. For example, a novel file format has been introduced to speed up reading of SWATH-MS maps using a 3D indexing approach. Another area of current research is the *in silico* annotation of SWATH-MS maps without the need of prior DDA reference libraries.

Typical Applications of SWATH-MS Similar to SRM, quantitative data is acquired in the three major categories described above: microbial, eukaryotic model organism, and clinical. As SWATH-MS was established only recently, most publications thus far are of a technical nature demonstrating the feasibility of the methodology. For example, besides quantifying the peptidome originating from whole cell lysates, SWATH-MS is also suitable for quantification of AP-MS experiments, as demonstrated in the case of using 14-3-3β protein as bait and quantifying changes in proteins associated with 14-3-3β as a function of stimulating the insulin-PI3K pathway in cultured human cells. An example of microbial study is the proteomic reorganization *Mycobacterium tuberculosis* undergoes during exponential growth, hypoxia-induced dormancy, and resuscitation. Similarly, yeast has been used to quantify proteome changes as a function of osmotic stress with high reproducibility between biological and

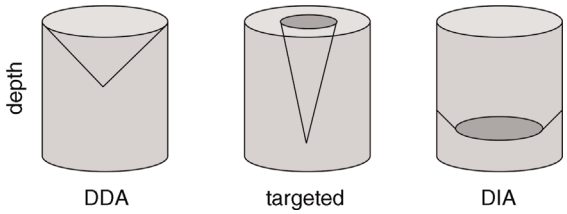


Figure 39.19 Summary of proteome coverage: the proteome is very complex due to post-transcriptional splicing of mRNA and post-translational modifications (PTMs) of proteins. DDA is a valuable method for identifying proteins and their PTMs, but the method suffers from the stochastic sampling approach and is not the most sensitive LC-MS/MS method. Targeted proteomics is the "gold standard" of quantitative proteomics and the most sensitive LC-MS/MS method, but covers only a small portion of the entire proteome. DIA is not as sensitive as Targeted, but all peptides identified are quantified on the MS2 level.

technical replica. An example of clinical application is the quantification of N-glycosylated serum peptidome of twins. Researchers conducting this study concluded that clinical serum biomarkers should be calibrated against genetic background and age adjusted.

39.5.9 Summary

Data dependent analysis of the proteome samples a wide range of the proteome to identify many peptides and therefore proteins. SRM consistently quantifies a large number of peptides across many samples and is the most sensitive LC-MS/MS method currently available. SWATH-MS is able to quantify a very large fraction of the expressed proteome (Figure 39.19).

39.5.10 Extensions

Besides DDA, SRM, and SWATH-MS, other bottom-up proteomics approaches have been developed or are currently implemented. Some methods are optimized for a particular mass spectrometer, but what unifies all current targeted bottom-up proteomics approaches is the goal of consistently quantifying as many peptides as possible across a biologically relevant experiment for a complete data matrix.

PRM Parallel reaction monitoring (PRM) is a targeted proteomics method that substitutes the third quadrupole in a typical QQQ setup with a high-resolution mass analyzer (e.g., an Orbitrap). This allows PRM to obtain a complete recording of all product ions produced by a specific precursor (instead of a pre-selected number as in SRM). In addition, the high mass resolution (parts per million compared to 0.2–1 *m/z*) in conjunction with the ability to monitor all ionizable product ions allows for higher specificity in PRM compared to traditional SRM. Specialized software is then used to extract individual fragment ions from all high resolution MS^2 or MS/MS spectra associated with a given precursor in order to reconstruct product ions traces. Note that while the product ions traces produced by PRM look similar to SRM traces, PRM records substantially more data and allows computational re-analysis, for example, using a different set of transitions if an interference is detected in the initial set of transitions. Compared to SRM, PRM allows for reduced assay development time.

PAcIFIC While the DDA method matured, the need for data independent acquisition was recognized. An early implementation of DIA was "precursor acquisition independent from ion count" (PAcIFIC) developed in the Goodlett laboratory. Using LC-MS/MS instrumentation of the time, a precursor isolation width window of *m/z* 2.5 was selected for MS^2 fragmentation and MS/MS spectra recorded (e.g., *m/z* 400–402.5). Next, the isolation window was shifted by 1.5 *m/z* and all precursors within the window of *m/z* 401.5–404 were selected for MS^2 fragmentation. The cycle ended when the window of *m/z* 411–413.5 was reached and was repeated throughout the LC gradient. In the next LC-MS/MS experiment, another isolation window was covered until the entire *m/z* range from *m/z* 400 to 1400 was completed. PAcIFIC required 67 LC-MS/MS experiments (or 4.2 days) to complete the entire mass range. Although no special software was required to annotate the MS/MS spectra, as the isolation width per precursor was still quite small, the PAcIFIC method was not implemented widely. But, PAcIFIC was one of the first attempts at DIA and changes to the precursor selection (now using a quadrupole instead of an ion-trap) make it now possible to perform SWATH-MS-like experiments on an Orbitrap

mass analyzer, together with a higher sensitivity and increased scan speed of the mass spectrometer.

MSX MSX is a targeted proteomics strategy that works similarly to SWATH but includes a de-multiplexing strategy for the highly multiplexed fragment ion spectra produced in DIA. In MSX, multiple precursors from non-consecutive isolation windows (e.g., five) are co-fragmented in each cycle and co-fragmenting different precursors in each cycle together allows for de-convolution of the different isolation windows during post-processing. In this manner a *m/z* 20 isolation window can be deconvoluted into five "virtual" isolation windows that have an effective width of only *m/z*, 4 allowing for much higher specificity. In this manner, the MSX approach has the potential to combine the throughput of SWATH-MS with the specificity of PRM.

MS^E A vendor specific implementation of DIA is MS^E, which requires both a specialized LC separation system and a special mass spectrometer. To minimize the number of co-eluting peptides, an LC system performing ultra high pressure called UPLC is required. Upon ionization, the precursors enter the special mass spectrometer, which is equipped with a collision cell that cycles between two states: low and high collisional energy (CE). At low CE, the precursor ions travel through the collision cell to be detected by the time-of-flight (ToF) mass analyzer (= MS^1). At high CE, fragmentation of all precursors occurs and the *m/z* of all detectable fragment ions is recorded by the ToF (= MS^2). Vendor specific software then deconvolutes the complex data for raw data visualization, identification, and quantification of peptides.

39.6 Stable Isotope Labeling in Quantitative Proteomics

Friedrich Lottspeich

Peter-Dörfler-Straße 4a, 82131 Stockdorf, Germany

39.6.1 Stable Isotope Label in Top-Down Proteomics

Unfortunately, mass spectrometry is not a quantitative method *per se*. An exact quantitative determination of analyte concentrations by MS can only be done when the mass spectrometric signal from a certain molecule is compared with the signal of a known concentration of a standard sample. This is almost impossible with proteome analysis, since very few pure standard proteins are available and, additionally, the measured signal intensities depend on many factors, including the sample matrix and the measurement parameters. Thus, to create quantitative mass spectrometric standard curves for all proteins of a proteome would not be possible for time and cost reasons.

At least for a relative quantification this problem can be overcome by so-called isotope labeling techniques. For a quantitative proteome analysis one state of the proteomes will be marked with "heavy" stable isotopes (^{13}C, ^{15}N, ^{2}H, ^{18}O) instead of the naturally occurring ("light") isotopes(^{12}C, ^{14}N, ^{1}H, ^{16}O). The so labeled proteins, and accordingly their proteolytic peptides, behave in principle in almost any separation process identically to the peptides with "light" isotopes. However, since the heavy and light forms of the label have a different molecular mass the modified proteins/peptides, although they have the same sequence, can easily be separated and distinguished by mass spectrometry. Thus, each protein/peptide may, due to its mass ("heavy" or "light"), easily be assigned to the corresponding "heavy" or "light" labeled proteome state.

It follows that after labeling the proteomes can be mixed and thus all further steps (e.g., separations and mass spectrometry analyses) can be performed under identical conditions.

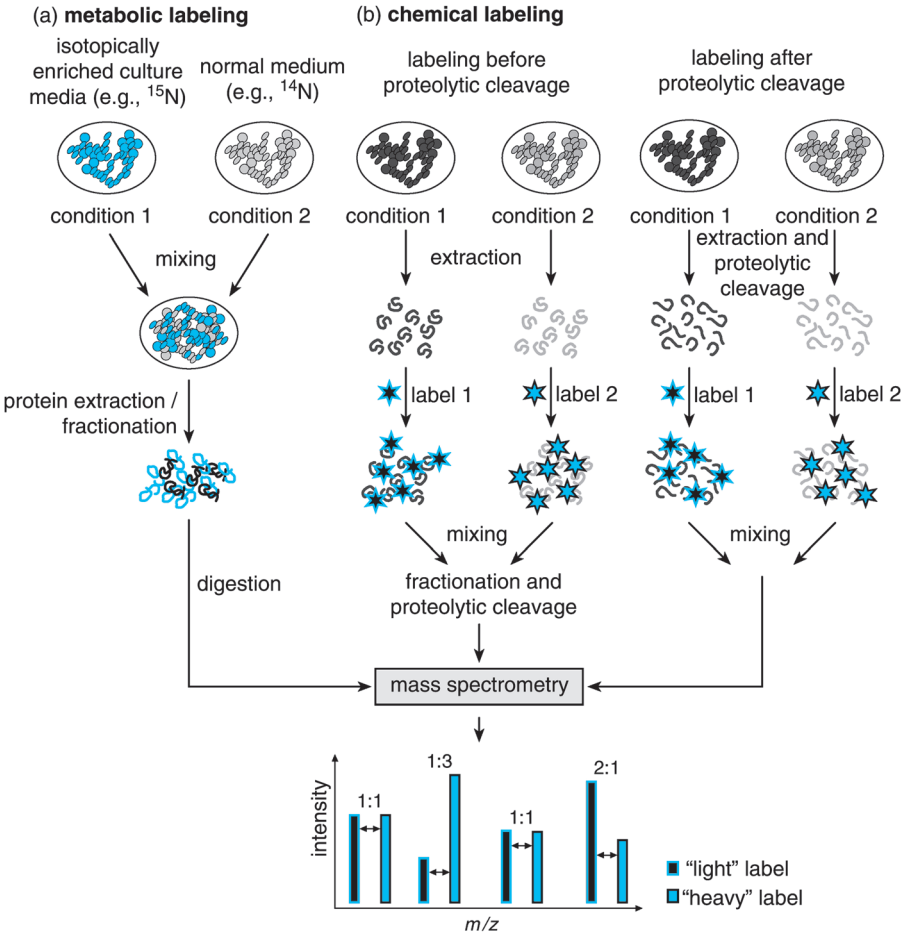


Figure 39.20 Schematic representation of the various methods of stable isotope labeling for relative quantitative analysis of proteins by mass spectrometry. (a) Metabolic labeling of proteins in cell culture with isotopically enriched culture media (e.g., as with ^{15}N isotope-labeled amino acids or salts) or with normal medium. For the relative quantitative mass spectrometric analysis the cells were mixed in a 1 : 1 ratio, and after cell disruption the proteins extracted, fractionated, and digested proteolytically. (b) Chemical labeling takes place after extraction of proteins either before or after proteolytic cleavage of proteins. Which amino acid is derivatized (i.e., the reaction specificity) is dependent on the reagent used. The mixing of the sample of state 1 and 2 takes place in each case after the introduction of the isotope-coded label. The proteolytic peptide mixtures are analyzed by mass spectrometry and the signal intensities of the peptides in their "light" and "heavy" form evaluated quantitatively.

A protein of the labeled "heavy" state (e.g., with ^{15}N isotopes) acts effectively as an internal standard for the corresponding protein from the other state (with ^{14}N-isotopes). When pooled, the relative ratio of a protein species from differently labeled states remains fixed during all further fractionations and analysis steps, since the inevitable loss when using multi-dimensional separation methods (e.g., chromatography, electrophoreses) always affects both protein species to the same extent. The ratio of the signal intensities of "light" and "heavy" forms of proteolytic peptides in the mass spectrum allows the relative quantification of their corresponding proteins in the starting sample. Thus, the isotope labeling techniques allow mass spectrometry to be used as a quantitative method. In addition, absolute quantification (sometimes required, for example, in systems biology) is possible by adding isotope labeled standard peptides/proteins to a proteome state.

The different methods of stable isotope labeling can be divided into two main classes: metabolic (*in vivo*) mark and chemical (*in vitro*) mark. Figure 39.20 shows the different strategies of the stable isotope labeling methods for relative quantitative analysis of proteins by mass spectrometry (for explanations see the text below). The essential difference between the individual techniques is the time point at which the proteome is labeled. When the isotope labeling is introduced at the level of the organism, cell, or proteins, one speaks of *top-down* proteomics strategies. If the proteome is cleaved prior to isotope labeling in peptides, proteome analysis is restricted to peptide-based *bottom-up* approaches (Section 39.5).

Metabolic Labeling In metabolic labeling stable isotopes (e.g., ^{14}N) are incorporated during the growth of the organism (Figure 39.20a). The great advantage of this approach is that the differently labeled samples from different proteomic states can be pooled immediately after

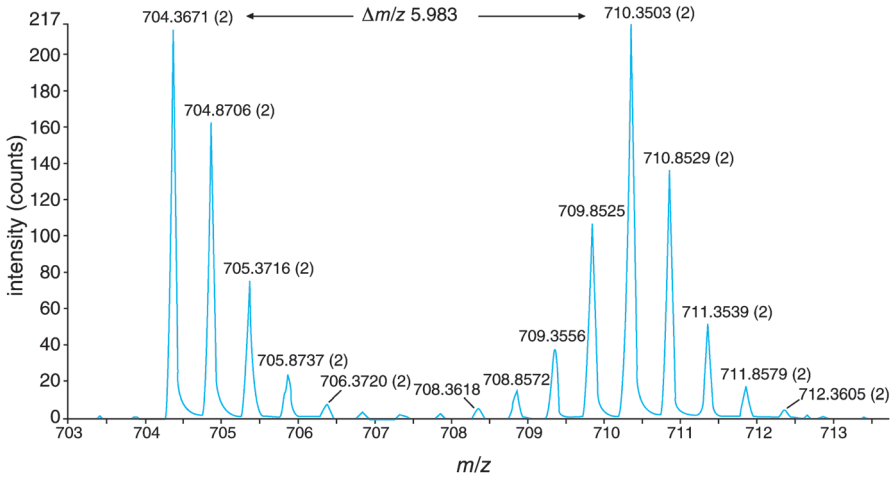


Figure 39.21 Detail of an ESI-MS spectrum of the metabolically labeled, with $^{14}N/^{15}N$ isotopes, peptide with the amino acid sequence LTYYYPEYETK. In the mass spectrum of the double charge state the "light" and "heavy" form of the peptide can be seen by the isotope intervals of *m/z* 0.5. A full introduction of ^{15}N atoms into the peptide sequence gives rise to a theoretical mass displacement of 11.964 Da. This matches perfectly with the observed mass difference of the doubly charged peptides ($z = 2$) of *m/z* 5983. Analysis of the signal intensities of $^{14}N/^{15}N$-peptide pair gives a ratio of 1 : 1.

cell harvesting and can be further processed together. Errors resulting from independent parallel sample preparation are minimized. Metabolic labeling methods are particularly suited for cell culture systems.

In the late 1990s the first trials for metabolic labeling of cell cultures were reported with stable isotopes using ^{15}N enriched culture media. During cell culture, into all proteins ^{15}N atoms are introduced instead of the naturally occurring ^{14}N isotope. The number of ^{15}N atoms introduced and therefore the mass shift compared to the naturally ^{14}N containing form depends on the amino acid sequence and no uniform peptide mass shifts are observed in the MS analysis. This complicates the relative quantitative analysis of mass spectrometric data and assumes that the primary structure of the protein/peptide under examination is known and appropriate software evaluation programs are available. Figure 39.21 shows a section of an ESI-MS spectrum of the doubly charged $^{14}N/^{15}N$-peptide pair with the amino acid sequence LTYYTPEYETK. A full introduction of ^{15}N atoms in the peptide sequence results in a theoretical mass shift of 11.964. This matches the mass displacement of $\Delta m/z$ of 5.983 in the observed spectrum (taking into account the two-fold charge state of the peptide pair). The relative signal intensities of the peptide in its "light" and "heavy" form display a 1 : 1 ratio in this example. The wider distribution of isotopes of the "heavy" peptide displayed in the mass spectrum results from the only 95% efficiency of the ^{15}N labeling. The relative signal intensities of the isotopes that occur in front of the monoisotopic peak at *m/z* 710.350 in the spectrum were caused by the statistical distribution of ^{14}N atoms (5%) and ^{15}N atoms (95%) in the primary sequence of the peptide (Section 14.1.4). Metabolic labeling with the ^{15}N isotope has also been applied to higher organisms such as *Caenorhabditis elegans*.

SILAC Another very elegant *in situ* method is stable isotope labeling with ^{13}C-, ^{2}H-, or ^{15}N-modified amino acids, which appeared in the literature as SILAC (stable isotope labeling by amino acids in cell culture). Originally, the introduction of isotopically labeled amino acids was used during cell culture to ensure the reliability of protein identification by MALDI-peptide mass fingerprint (PMF) in combination with database searching. This concept has since been extended in the laboratory of M. Mann and used for MS-based relative quantitative protein analysis.

In one experiment, two SILAC-cell samples are cultured in parallel in a culture medium with an unchanged amino acid composition, the other in a medium which contains one or two isotopically substituted amino acids. In order to achieve a complete introduction of these amino acids in the proteins, at least five cell cycles have to be completed. After cell harvesting, the samples can be directly pooled in a 1 : 1 ratio, and relative to each other quantitatively analyzed by MS.

For the MS analysis with SILAC, usually $^{13}C_6$-arginine, $^{13}C_6$-lysine, or $^{13}C_6$, $^{15}N_4$-arginine and $^{13}C_6$, $^{15}N_2$-lysine are used as modified amino acids. These are favorable since in most protein studies trypsin is used as the proteolytic enzyme, which cuts C-terminal arginine and

lysine. If the tryptic cleavage of a protein sample is specific and complete, the isotope coded arginine and lysine residues are located exclusively C-terminal to the cleavage products. This results in shifts in the mass spectrum masses of 6 Da for $^{12}C_6/^{13}C_6$-arginine or lysine peptide pairs, 10 Da for $^{12}C_6$, $^{14}N_4/^{13}C_6$, $^{15}N_4$-arginine peptide pairs, or 8 Da for $^{12}C_6$, $^{14}N_2/^{13}C_6$, $^{15}N_2$-lysine peptide pairs. Apart from the relative quantitative statement from the ratio of signal intensities of peptides in their "light" and "heavy" form, sequence-specific information from the MS^2 spectra is also obtained. A general disadvantage of labeling with deuterated compounds (e.g., $^2H_4$-lysine) is that ^1H- and ^2H-labeled analytes have slightly different chromatographic properties (isotope shift) and thus may complicate the relative quantification in an, for example, HPLC/ESI-MS system. For the quantitative automatic evaluation of SILAC experiments, high-performance data processing software (e.g., MaxQuant) is available.

By further SILAC-amino acid combinations a triple labeling may also be performed. In addition to the relative quantitative study of the expression of proteins, a large number of new applications for SILAC can be found. SILAC in combination with high resolution mass spectrometry can monitor and time-resolve biological signaling processes, or identify *in vivo* methylation sites in proteins and protein–protein interactions. In newer MS-based quantitative proteomics the SILAC technique was successfully used to label primary cell cultures or the proteomes of higher organisms such as *Drosophila melanogaster* or *Mus musculus*. For the labeling of SILAC yeast proteomes it is noted that yeast strains must be generated, which are auxotrophic for arginine and lysine, in order to ensure complete incorporation (>98%) of the isotope labeled amino acids. The proteome of autotrophic organisms may be only incomplete and difficult to mark with SILAC. Therefore, for MS-based quantitative proteomics studies of green plants, such as *Arabidopsis thaliana*, the metabolic $^{14}N/^{15}N$-labeling or chemical labeling methods may be used. At this point it must be pointed out that for a quantitative analysis under two different experimental conditions, the tests must be carried out independently several times in order to draw conclusions about the technical and biological variance of the data. This applies equally well to the methods described below.

Chemical Stable Isotope Labeling Chemical stable isotope labeling is a universal method, since it can be applied to almost any sample, such as cultured cells, body fluids, or tissue biopsies from any organism. The introduction of stable isotopes by a suitable chemical reaction can be made directly in the intact proteins and is so designed for the top-down proteomics strategy (see above, Metabolic Labeling). These reagents, however, are also generally suitable for labeling of proteolytic cleavage products after enzymatic digestion of the proteins for the bottom-up proteome strategies (Section 39.5).

Among the strategies currently available are the cysteine specific labeling with isotope-labeled affinity tags (isotope-coded affinity tags, ICATs) or amino group specific labeling of proteins with ICPL (isotope-coded protein label).

ICAT – Isotope-Coded Affinity Tag Method Initial results achieved with such an analysis strategy were published by Aebersold already in 1999. Figure 39.22 shows the ICAT reagent and Figure 39.23 shows the principle of the method. All proteins of a proteome state are derivatized with the ICAT reagent in its "light" version on the SH groups of the cysteine residue. The proteins of other proteome state are labeled with the "heavy" version of the ICAT reagent. The two reagents are chemically identical, but differ in terms of eight hydrogen atoms replaced by deuterium atoms.

Therefore, after derivatization a protein molecule that, for example, has a cysteine residue, from the first ("light") proteome state has a molecular weight that is lower by 8 Da than the same

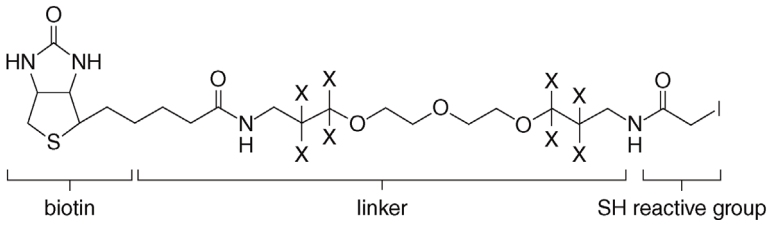


Figure 39.22 The ICAT reagent, X = hydrogen in light reagent, deuterium in heavy reagent.

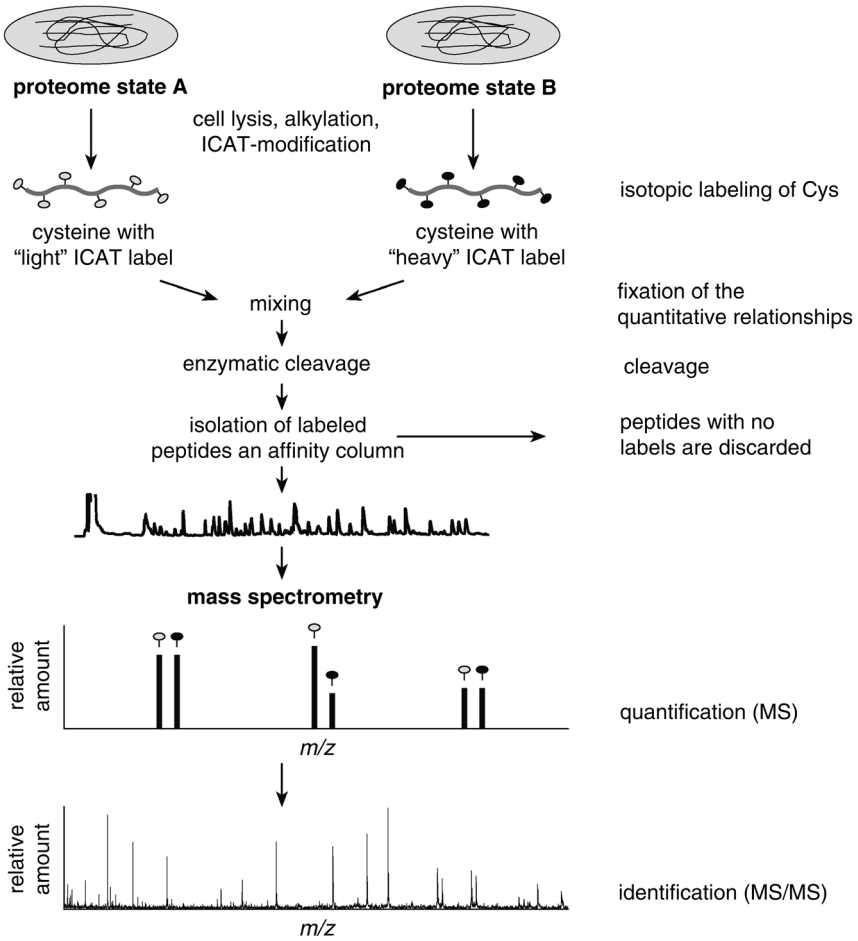


Figure 39.23 Schematic representation of the ICAT technique. The cysteine residues of all proteins in the control (e.g., healthy tissue) are reacted with the light ICAT label, the cysteines of the condition to be tested (e.g., diseased tissue) with the heavy label. The differently labeled proteomes are mixed in a 1 : 1 ratio, enzymatically digested, and the ICAT-labeled cysteine containing peptides are isolated over an affinity column and subsequently identified by LC/MS/MS and quantified relative to each other.

protein from the second ("heavy") state. The two isotopically labeled proteome states are mixed in a 1 : 1 ratio and enzymatically cleaved into peptides. Since only the cysteine-containing peptides are modified, only such peptides are present in the light and heavy form. Thus, only those peptide pairs can be assigned to the different proteome states. Because the ICAT reagent carries an affinity tag (biotin), all modified peptides can be isolated over a streptavidin affinity column. After optional further separation steps the co-eluting ICAT peptide pairs are analyzed by mass spectrometry. Using special software programs the peptide pairs that differ by 8 Da (or a multiple of 8 if more than one cysteine is present in the peptide) are quantified and identified on the basis of MS^2) data. The signal intensities or peak areas of the peptide pairs in the MS spectra reflect the relative amount of the peptides and therefore also their corresponding proteins in the individual proteome states.

The basic weakness of the ICAT method is that the isotopic labeling takes place on a rare amino acid. After derivatization and proteolytic cleavage of proteins only the cysteine-containing peptides are isolated by affinity chromatography. This means that generally very low sequence coverage is achieved and only those proteins can be quantitated that contain cysteine in the primary sequence. Protein isoforms, degradation products, or post-translational modifications that are not located in the cysteine-containing peptide are not recognized. In addition, the ICAT reagent is relatively large, does not react completely, and shows unspecific reactions caused by rather long reaction times. In addition, thioether is chemically labile with atmospheric oxygen, which may lead to uncontrolled removal of the label by β–elimination; the relative quantification of a protein is usually based only on a few peptides. There are today a new generation of ICAT reagents with $^{12}C/^{13}C$ isotope and acid-labile linkers available. An interesting application of these ICAT reagents is, for example, the quantitative study of oxidative thiol modifications in proteins.

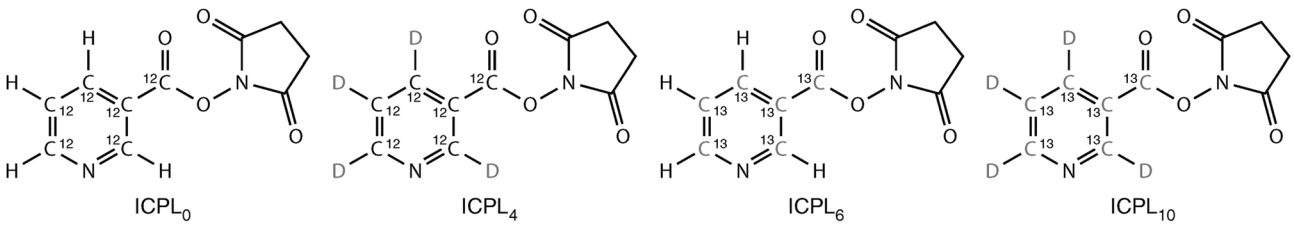


Figure 39.24 The four ICPL reagents. The heavy stable isotopes ^{13}C and ^{2}H (D, deuterium) are shown in gray.

ICPL – Isotope Coded Protein Label Another protein-based proteome analysis that employs isotopes is the ICPL method. For a comparative proteomic analysis the proteins of different proteome states are fully marked on the numerous existing amino groups with the different ICPL isotopologues (Figure 39.24). Thus, after cleavage all lysine-containing peptides of a protein can be used for the quantification. The ICPL reagents are currently available in four different isotopic variants. They change the properties of the proteins, but still allow fractionation with all separation techniques established in protein chemistry. The operation of the method is depicted in Figure 39.25. After labeling, the proteomes labeled with the different isotopic reagents are combined and, thus, the relative proportions of the individual proteins of the different states are

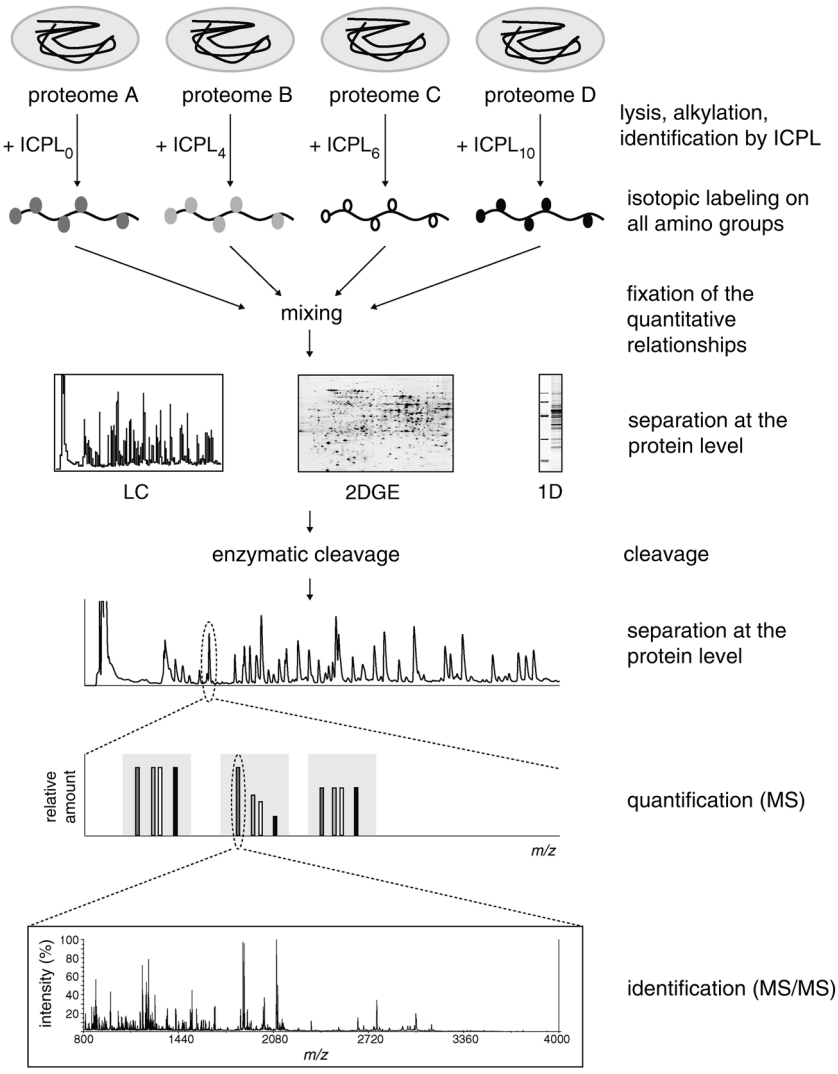


Figure 39.25 All proteins of the proteomes under investigation are labeled on all amino groups with the various ICPL reagents. The labeled proteomes are combined and separation at the protein level is carried out as efficiently as possible (preferably also multidimensionally). The proteins of the now relatively low complex fractions are enzymatically digested and the peptides (eventually after further separations) are analyzed by mass spectrometry, quantified, and identified.

fixed. To reduce the complexity at the protein level both electrophoretic techniques (isoelectric focusing, 1D-PAGE, 2D-PAGE) and chromatographic separation methods or combinations thereof can be used. Information on the relative proportions of the corresponding proteins from the various states is also retained during multidimensional fractionation processes. The strategic goal of this step is to separate the proteome in a large number of factions, each of which contains only a small number of proteins. The proteins in the individual fractions have then to be cleaved enzymatically into peptides. When using the ICPL approach a tryptic cleavage is arginine-specific because all lysine residues are derivatized and no longer accessible to cleavage. Double cleavage with two proteases (e.g., trypsin and Glu-C) is recommended to obtain smaller peptides, which are easier to analyze by mass spectrometry. After cleavage, relatively simple peptide mixtures are obtained that, since there should be only a few proteins present, are simple and can be analyzed by mass spectrometry and quantified automatically. Dedicated software programs like ICPLQuant, MaxQuant, or ProteinScape (Bruker) can handle the multiplexed spectra of the ICPL workflow and recognize the peptide pairs. The analysis on the MS^1 level already indicates which peptides/proteins differ quantitatively in different proteome states. Only those peptides must be further analyzed by MS^2 to identify them and deduce the corresponding proteins. Protein isoforms, degradation products, and post-translationally modified proteins can be separated at the protein level, in which case each of these protein species is recorded and quantified separately. The ICPL technology enables a comprehensive, quantitative proteomics analysis from various samples (body fluids, tissues, etc.) and can create fast and efficient differential protein patterns where the quantification is obtained at the MS^1 level.

Chromatographic Separation Methods, Chapter 10

Electrophoretic Techniques, Chapter 11

Cleavage of Proteins, Chapter 9

Mass Spectrometry, Chapter 15

39.6.2 Stable Isotope Labeling in Bottom-Up Proteomics

To circumvent the quantification problems with the label-free techniques isotope labeling techniques can also be used with bottom-up proteome strategies (see Section 39.5.5, 39.5.7 and Figure 39.20).

Non-isobaric Labeling With non-isobaric labels the peptides produced differ in molecular mass. Therefore, the quantification can be performed on the MS^1 level.

^{18}O Labeling Chemical labeling with ^{18}O is achieved by an enzymatically (e.g., trypsin) catalyzed reaction with $H_2{}^{18}O$ directly in the cleavage reaction. Two ^{18}O atoms are incorporated in the C-terminal carboxyl group, thus producing a +4 Da shift that can easily be detected in MS^1 analysis.

The mass shift from ^{18}O incorporation does not alter the chromatographic separation or the ionization efficiency of the labeled peptides. Loss of the isotopic label has been reported since the mechanism of ^{18}O introduction is reversible. Therefore, immobilized trypsin should be used; the trypsin can be removed from the sample after the labeling step.

Labels with Different Numbers of Stable Isotope Atoms (^{13}C or ^{2}H) in the Reagent (e.g., ICPL, see above) Mostly amino group specific reagents that react with all ϵ-amino group of lysines and the amino terminal amino groups are used to introduce the labels, usually containing ^{13}C or ^{2}H isotopes. Predominantly, as reactive group hydroxysuccinimide esters are used and most of the peptides in a tryptic or LysC digest are labeled. Here the isotopic peptide pairs may also be identified and quantified already in the MS^1 analysis. Higher multiplexing is possible; however, the complexity of the spectrum increases accordingly.

After inserting the marker the various proteome states can be pooled, similarly to the top-down proteomics (see section above on Metabolic Labeling). By evaluating the introduced mass difference the assignment of the peptides to the individual proteome states is ensured. Although theoretically almost all peptides should be marked and high sequence coverage in the mass spectrometric analysis seems attainable, it can be seen in practice that such expectations are not met. This is mainly due to the enormous complexity that is generated by the enzymatic protein cleavage. Tens of thousands of proteins are cleaved to give several hundred thousand peptides. In addition, by marking the different weights of reagents the complexity of the peptide

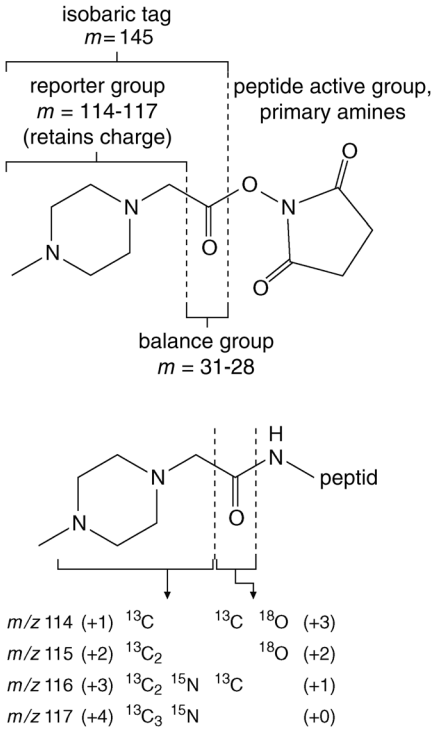


Figure 39.26 Four isobaric iTRAQ reagents (Sciex/Sigma-Aldrich).

mixture is further multiplied (e.g., quadrupled for an experiment with four different isotope reagents). This complexity exceeds the capacity of today's chromatographic or electrophoretic methods, so that only a portion of the proteome can be quantified by the subsequent analysis of such mixtures by high-resolution mass spectrometry. Therefore, multidimensional peptide separations are usually necessary for a comprehensive peptide-based proteome. The development of modern mass spectrometers and the corresponding software in recent years is tailored to this peptide-based workflow. As a consequence, today, an automatic, sensitive, and quantitative high-throughput analysis of peptides, partially including mass spectrometric identification, is routine.

Isobaric Labeling The currently most attractive peptide-based approaches based on isotopic labeling are carried out with isobaric reagents, like iTRAQ (isobaric tags for relative and absolute quantitation) (Sciex/Sigma-Aldrich) (Figure 39.26) or TMT (tandem mass tag) (Thermo Scientific). Here in the various proteome states isobaric labels can be introduced after enzymatic cleavage onto amino groups. The corresponding peptides of a protein from the various proteome states have the same mass. Only in the MS/MS analysis are reporter groups released that are specific for each reagent and the quantitative relations for the different proteome states can be recognized (Figure 39.27).

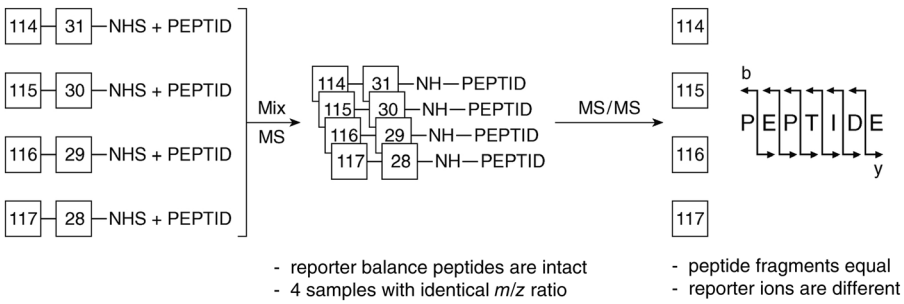


Figure 39.27 In the iTRAQ method four enzymatically cleaved proteome states are labeled with one of the isotope iTRAQ labeled reagents. The derivatized peptides from all states are isobaric and co-elute (e.g., in a chromatographic separation). In the MS/MS analysis of such a peptide mixture, the four reporter groups (114–117) are released and reflect in their relative intensity to each other the relative amounts of peptide in the single proteome states. Source: Ross, P. L. et al. (2004) *Mol. Cell. Prot.*, 3, 1154-1169. With permission, Copyright © 2004, by the American Society for Biochemistry and Molecular Biology.

Some major advantages of this method are:

- a high intensity of the peptide signal MS spectra, since all the signals of a certain peptide from all states co-migrate, which usually allows a high quality MS/MS spectra of the peptide for identification;
- up to ten different forms of an isobaric reagent are commercially available, from which it follows that up to ten different samples can be analyzed in a single experiment – this allows increased throughput associated with reduction of costs;
- despite multiplexing there is no increase in MS^1 complexity.

One has to keep in mind that the methods using isobaric labels in combination with the bottom-up strategy have all the drawbacks of peptide-based approaches as given above (Section 39.5.2 and Chapter 14, Introduction). In addition, all peptides must be analyzed with MS/MS techniques because the quantification is carried out by releasing the reporter ions and thus the quantitative relationships are only recognizable in the fragment spectra. Therefore, all peptides, including those that do not change in their amount and are often not of interest, have to be analyzed by using MS/MS. However, this is done quickly and automatically with the latest generation of mass spectrometer. Owing to the high complexity of the spectra, there is a risk that the MS^1 spectrum includes not only the signals of a single peptide. During MS^2 fragmentation all peptides liberate the reporter ion, but not all the peptides are identified in the MS/MS analysis, which in the end may give false quantifications.

Further Reading

Sections 39.1–39.3 and 39.6

Kellermann, J. and Lottspeich, F. (2012) Isotope-coded protein label. *Methods Mol. Biol.*, **893**, 143–153.
Liebler, D.C. (ed.) (2002) *Introduction to Proteomics – Tools for the New Biology*, Humana Press, Totowa.
Lottspeich, F. and Kellermann, J. (2011) ICPL labeling strategies for proteome research. *Methods Mol Biol.*, **753**, 55–64.
Posch, A. (ed.) (2008) *2D PAGE: Sample Preparation and Fractionation*, vol. 1 + 2, Springer Verlag, Berlin.
Vogt, A., Fuerholzner, B., Kinkl, N., Boldt, K., and Ueffing, M. (2013) Isotope coded protein labeling coupled immunoprecipitation (ICPL-IP): a novel approach for quantitative protein complex analysis from native tissue. *Mol. Cell. Proteomics*, **12**, 1395–1406.
Von Hagen, J. (ed.) *Proteomics Sample Preparation*, Wiley-VCH Verlag GmbH, Weinheim.
Walker, J.M. (2002) *Protein Protocols Handbook, Methods in Molecular Biology*, Springer Verlag, Berlin.
Wilkins, M.R., Williams, K.L., Appel, R.D., and Hochstrasser D.E. (eds) (1997) *Proteome Research: New Frontiers in Functional Genomics*, Springer, Heidelberg.

Section 39.4

Chait, B.T. (2006) Chemistry mass spectrometry: bottom-up or top-down? *Science*, **314** (5796), 65–66.
Han, X., Jin, M., Breuker, K., and McLafferty, F.W. (2006) Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. *Science*, **314** (5796), 109–112.
Kelleher, N.L. (2014) A cell-based approach to the human proteome project. *J. Am. Soc. Mass Spectrom.*, **23**, 1617–1624.
Kelleher, N.L. (2004) Top-down proteomics. *Anal Chem*, **76** (11), 197A–203A.
Reid, G.E. and McLuckey, S.A. (2002) 'Top down' protein characterization via tandem mass spectrometry. *J. Mass Spectrom.*, **37** (7), 663–675.

Section 39.5

Barnidge, D.R., Dratz, E.A., Martin, T., Bonilla, L.E., Moran, L.B., and Lindall, A. (2003) Absolute quantification of the G protein-coupled receptor rhodopsin by LC/MS/MS using proteolysis product peptides and synthetic peptide standards. *Anal. Chem.*, **75**, 445–451. First SRM proteomics experiment.
Desiderio, D.M. and Kai, M. (1983) Preparation of stable isotope-incorporated peptide internal standards for field desorption mass spectrometry quantification of peptides in biologic tissue. *Biomed. Mass Spectrom.*, **10**, 471–479. First stable isotope labeled reference peptides for robust sample comparison.

Ebhardt, H.A., Root, A., Sander, C., and Aebersold, R. (2015) Applications of targeted proteomics in systems biology and translational medicine. *Proteom. Rev.*, **15** (18), 3193–3208. SRM literature.

Liu, Y., Buil, A., Collins, B.C., Gillet, L.C., Blum, L.C., Cheng, L.Y., Vitek, O., Mouritsen, J., Lachance, G., Spector, T.D., Dermitzakis, E.T., and Aebersold, R. (2015) Quantitative variability of 342 plasma proteins in a human twin population. *Mol. Syst. Biol.*, **11**, 786. Application of SWATH-MS for longitudinal twin study.

Picotti, P., Bodenmiller, B., Mueller, L.N., Domon, B., and Aebersold, R. (2009) Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell*, **138**, 795–806. First perturbed protein network quantification using SRM.

Metabolomics and Peptidomics

40

Peter Schulz-Knappe and Hans-Dieter Zucht

Protagen AG, Otto-Hahn-Str 15, 44227 Dortmund, Germany

Over the last decade bioanalytical sciences have evolved from firmly established areas such as genomics to a whole set of novel "omics"–type of technologies. These disciplines are often defined to investigate comprehensively a certain class of biomolecules. For example, whilst transcriptomics covers the function and expression of messenger-RNA-molecules, proteomics and peptidomics have emerged as focusing on the comprehensive evaluation of proteins and biologically processed peptides. The research area referred to as metabolomics is dedicated to the analysis of the multitude of low molecular, organic, non-polymeric molecules (metabolites) in order to assess biological phenotypes or biochemical states. Non-targeted metabolite profiling involves the identification and characterization of a large number of metabolites and their precursors. In contrast, targeted metabolite profiling focuses on quantitative changes in metabolites of interest (e.g., amino acids, carbohydrates, steroids, and fatty acids) based on a priori knowledge of the biological function or metabolic pathway.

The driving forces for innovation in metabolomics research are based on the improvements in analytical technology and instrumentation facilitating the quantitative and highly specific multiplex analysis of thousands of different molecules. A second factor is the revolution in information technology providing tools to handle large sources of measurement data. Hence bioinformatics has become an integral part of metabolomics.

The following terminology is used in this field, which splits it into certain aspects:

- *Metabolomics:* comprehensive analysis of all metabolites of a cell, an organism, or a body fluid under well-defined conditions.
- *Metabonomics:* analysis of biochemical alterations under the influence of a disease, a drug intervention, or toxin.
- *Metabolite profiling:* selective analysis of a certain subset of metabolites such as amino acids or fatty acids.
- *Metabolic profiling:* Comprehensive survey analysis of a large set of biomolecules often in a time dependent manner and quantitatively. Applications are the clinical and pharmaceutical analysis of a drug, its metabolites including its kinetics, conversion into intermediates, degradation/clearance; often used in conjunction with the term *metabolite profiling*.
- *Metabolic fingerprinting*: classification of samples with respect to biological relevance or origin of source without identification of individual metabolites, which involves rapid, high-throughput global analysis to discriminate between samples of different biological status or origin.
- *Metabolite target analysis*: description of relevance of a target by analyzing its substrate, for example the analysis of metabolites of a distinct enzymatic cascade that is altered by abiotic or biotic influences.

Metabolomics The comprehensive analysis of metabolites of a cell, an organism, or a body fluid under defined conditions.

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

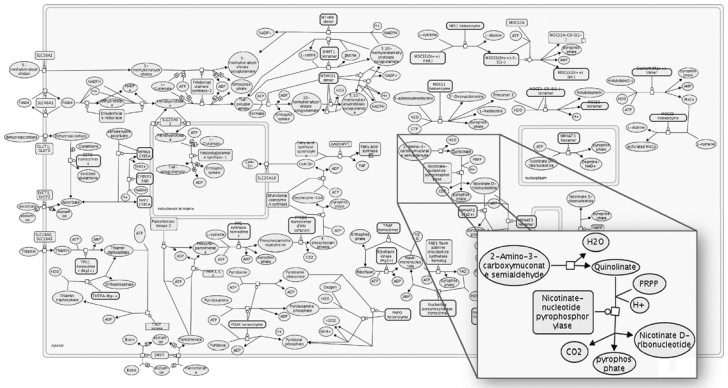


Figure 40.1 Representation of complex structures of biochemical pathways. Adapted and modified from "Metabolism of water-soluble vitamins and cofactors" (reactome.org; (Homo sapiens); Jassal B., D'Eustachio P., Stephan R.; 10.3180/REACT_11238.1).

Metabolome The multitude of metabolites in a biological system. The metabolic state can be described mathematically considering metabolic entities, their concentration, spatial distribution, and time dependent alterations.

The terminology listed here is somewhat ambiguous. Several definitions overlap in their meaning and their use throughout the scientific literature. Such a classification aims to outline different aspects used to analyze the metabolome.

The analysis of all "omics" classes has the goal of allowing for a comprehensive, encyclopedic picture of a biological system such as a cell, or an organ, and its actual state whilst influenced by environmental factors or drugs. The aim with "omics" approaches is to understand in a more holistic way how genetic information translates into function and provides a broad data basis, hopefully sufficient, for an interpretation of an emerging phenotype. Thus, in systems biology thinking, transcriptomics, proteomics, peptidomics, and metabolomics span the entire path from genotype towards a phenotype.

Historically, knowledge about metabolic synthesis and degradation processes in cells was initially dedicated to understanding the basic organizational principles characteristically mediated through enzymatic catalysis by enzymes. Milestone achievements have been the discovery of common basic organizational pathways for providing metabolic building blocks, handling energy and their regulation. The common pathways such as the *Krebs cycle*, the *pentose phosphate pathway, or glycolysis* are profound examples of such basic organizational principles. They illustrate a tight relationship between metabolites and their processing enzymes. The regulatory principles are mainly based on molecular interactions, and enzyme kinetics are able to describe individual steps. The identification of all the members of such pathways was the basis on which to create directed graphs mapping the multitude of interactions between the compounds depicted in textbook biology, as shown in Figure 40.1, as a kind of knowledge representation. Such textbook images provide a handle for a qualitative general understanding and are valuable educational tools today. However, they do not allow for a comprehensive representation of quantitative, qualitative, spatial, and temporal information types of modern "omics"-type approaches. This explains the important role of today's knowledge representation of metabolomics using computational methods. Today, the knowledge base is heavily supported with computational methods, such as dedicated databases, networking tools, and visualization software to interact with metabolomics databases. Examples are the KEGG (*Kyoto Encyclopedia of Genes and Genomes*) and Reactome (http://www.reactome.org) databases or Wikipedia-based systems like *Wiki pathways* (http://www.wikipathways.org/index.php/WikiPathways), which serve the metabolomics researcher.

Today, many organisms such as humans, bacteria, and some plants have been thoroughly studied, and the number of known metabolites can be estimated as quite complete. Surprisingly, the number of metabolites in humans appears to be quite small. Current estimates for humans are in the range of 2500 catalogued metabolites (excluding the gut microbial metabolites), whereas the number in plants exceeds 100 000. How can such a relatively small number of metabolites account for the known complex phenotypes? One possible explanation might be the highly dynamic changes in concentrations and the concerted, time- and location-dependent regulation of serial as well as parallel reactions. Highly complex regulations can appear even with a seemingly restricted repertoire of active players (= metabolites).

Interestingly, the number of protein and peptide species is quite large in humans. Estimates of the number of individual protein species in humans range from 100 000 to 1 000 000. The reason for the diversity of proteins lies in the fact that posttranslational modifications within

those polymers can create a large number of variants based on permutation. As peptides result from multiple, protease mediated cleavages of proteins (peptein = Greek for "digestion"), their number is estimated to be even higher, and several million different species might be present (Figure 40.2). Peptides as processing products of the respective precursor protein can be considered in some ways to be "metabolites" of proteins: their molecular mass ranges typically from 1 to 15 kDa (equivalent to chain lengths of 7–150 amino acids).

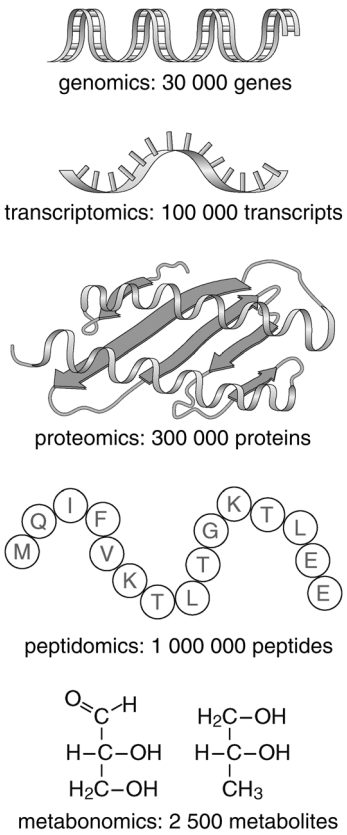


Figure 40.2 Organizational structure from genotype to phenotype.

40.1 Systems Biology and Metabolomics

The overarching goal of metabolomics is the detailed understanding of metabolic rules, integration, and regulations of entire organisms. Over recent years a novel research field has evolved that is recognized as *systems biology*. The "omics" disciplines complemented with systems theory form the integral parts of systems biology approaches. Systems biology describes metabolic systems in a qualitative, quantitative, and time- and space-resolved manner. Supporting disciplines are systems theory, mathematics, biostatistics, computing, and physics. Life can be defined as the coordinated and complex interaction of (mostly organic) chemical molecules in a living system in the context of the local and distant environment. Metabolomics delivers highly relevant data and knowledge about interactions of metabolites, equilibria, and substance fluxes as response to the environment (Figure 40.3). A common interpretation in system biology is that self-organizing biological systems exhibit distinct properties and reactivities. These are adapted to rapid and slow changes in the environment based on regulation and evolution. Systems biologists try to simulate systems based on computer simulations of metabolites and certain heuristic models. A growing number of genomes are now known (including many mammals, bacteria, viruses, and plants). Whereas a genome represents a static description of an organism, the other biological systems represent themselves in qualitatively as well as quantitatively highly dynamic, interdependent data. Importantly, external influences regularly influence many, if not most, parameters in a "dose-dependent" way, backwardly influencing gene expression and all the many downstream systems, especially proteins and peptides, and also sugars, lipids, and

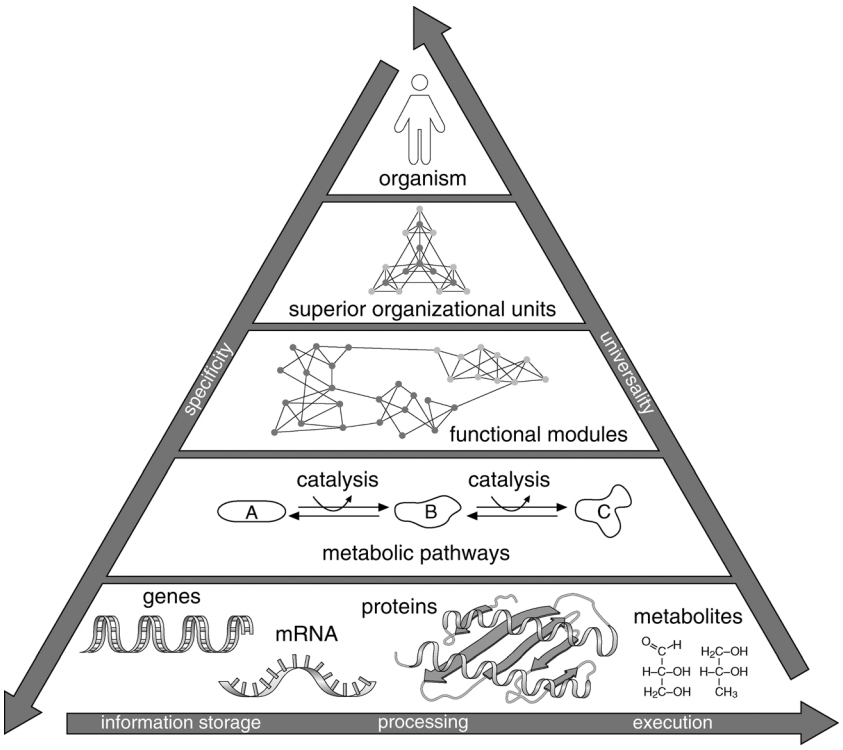


Figure 40.3 Hierarchical representation of the complexity of living systems. Genomics, transcriptomics, proteomics, peptidomics, and metabolomics generate the basis for the formulation of biochemical pathways. The organizational hierarchy over five layers is shown in this sketch adapted and modified from Oltvai, Z. N. and Barabasi, A.-L. (2002) *Science* **298**(25), 763.

metabolites. All this is necessary to allow for growth, differentiation, repair, and reproduction, at the right time and in the correct location.

40.2 Technological Platforms for Metabolomics

Metabolites are chemical entities that are – in contrast to genes, proteins, peptides, and polysaccharides – small, non-polymeric molecules. The listing below gives an overview of metabolites. Metabolites are analyzed using all the methods of conventional chemical analysis and analytical instrumentation. Highly resolving spectroscopy and spectrometry methods have been developed in past years that are able to exploit small specific characteristics of biomolecules. Especially useful and common are methods of molecular spectroscopy such as NMR, IR, and Raman spectroscopy, and spectrometric methods, especially mass spectrometry, which is often coupled to some kind of sample preparation. The coupling of analytical instruments to automated molecular separation technology such as high-performance liquid chromatography or gas chromatography enhances the resolution power and selectivity of an analytical workflow, providing procedures that can handle thousands of molecules in one analysis run. Most technologies need extraction procedures (e.g., MS, LC-MS, and GC-MS) and are restricted to snapshot analysis but some technologies (e.g., Raman and NMR) can even provide real time data of living organisms. The analytical platforms offer often both qualitative as well as quantitative data thus enabling cataloging of survey analysis and kinetics.

Metabolites:

- amino acids and their derivatives
- sugars
- ATP and other signaling molecules
- cholesterol and derivatives
- fatty acids, hydroxy acids, bicarbonic acids, polyamines

For a comprehensive analysis of metabolomes as a complex mixture of diverse components it is necessary to utilize a series of different technologies. To do this in a meaningful manner, it is necessary to distinguish methods of low and high specificity. Methods with low specificity usually address general features of a sample that are not based on individual molecules alone.

NMR, Chapter 18

Nuclear magnetic resonance (NMR) and infrared spectroscopy (IR) are most often used. These methods display chemical moieties of groups of metabolites. Results generated form a molecular fingerprint, and high numbers of samples can be analyzed in a short time. This is used for classification of samples, but in terms of sensitivity or structural identification of individual components these methods have clear limitations. Biofluids such as blood plasma, serum, urine, or cerebrospinal fluid can be assessed easily with NMR and IR, as sample preparation and measurement effort are very low. In NMR spectroscopy signals are generated from nuclei of certain atoms such as ^{1}H or ^{13}C. After excitation, they emit signals that alter their frequency and strength according to the molecular context within their molecular structure. This leads to defined alterations of nuclear magnetic resonance spectra. NMR spectroscopy yields many different signals from a single analyte, because organic biomolecules contain many functional groups with distinct signatures. When analyzing complex compound mixtures or whole organisms, organs, or body fluids such as blood plasma by NMR, the resulting spectra will be dominated by the main components present in the sample. An example from liver tissue extract is depicted in Figure 40.4. In a fairly simple way the difference between treated and untreated liver tissue is shown.

Infrared Spectroscopy, Section 7.4

The multitude of signals derived from a single biomolecule explains why only a limited number of analytes are readily accessible by NMR, and that minor components are frequently missed. The big advantage of the methodology lies in its ease of use, relatively low costs, simple sample preparation, and high-throughput capabilities, though instrument costs are still high. Without prior hypothesis a screening for differences of biological samples to reference samples is achieved. For this reason an established place for NMR is in areas such as food quality control, toxicology, and drug abuse. NMR-spectroscopy is an ideal tool for classification of biological samples as a non-destructive test procedure with the ability to analyze fully functional, living organisms.

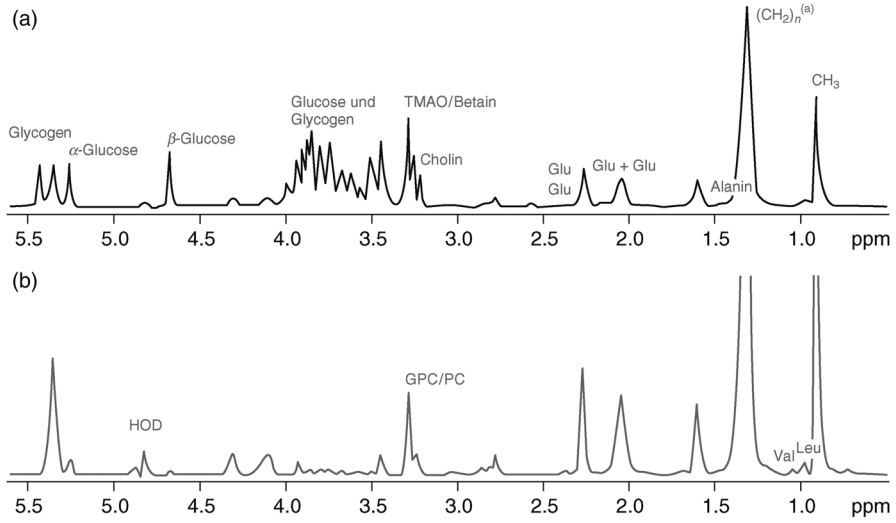


Figure 40.4 Differential ^{1}H NMR spectroscopic analysis of liver tissue. Differential comparison of an untreated control (a) to a sample retrieved after application of a liver-active compound (b). Adapted and modified from Coen, M. and Kuchel, P. (2004) *Chem. Aust.*, **71** (6), 13–17.

40.3 Metabolomic *Profiling*

Higher specificity in bioanalytical sciences is regularly achieved using high-resolution technologies such as mass spectrometry. Resolution, sensitivity, and selectivity are significantly modified and enhanced if the analysis methods are complemented by separation technologies such as gas chromatography (GC), high-performance liquid chromatography (HPLC), or capillary zone electrophoresis (CE) (Figure 40.5). In coupling mass spectrometry to such separation technologies the retrieval of quantitative and qualitative data can be achieved for individual molecular species, even if the starting sample consists of a very complex mixture. This strategic approach is referred to as "profiling". Depending on the class of analytes to be investigated the analysis methods can be prioritized. The necessary detail of analysis determines which technology or technology combination is selected. The resulting data, whether from metabolic, proteomic, or peptidomic profiling, have a high degree of similarity, as they represent a qualitative and (semi-) quantitative description of a large number of components. In an ideal scenario, identification of analytes is achieved simultaneously with quantification.

Mass Spectrometry, Chapter 15

Chromatographic Separation Methods, Chapter 10

Capillary Electrophoresis, Chapter 12

In general, analyzing any of the above-mentioned substance classes, one relevant requirement is the unbiased, robust, and reproducible extraction of class members from their respective biological matrix. This is even more relevant in all cases where high-resolution analytic technologies are utilized. Especially if in addition to the presence the relative or absolute concentration of many analytes is also desired the careful selection and validation of sample collection, storage, and sample preparation are mandatory to avoid selective enrichment/depletion of compounds, which would lead to strong bias in retrieved data. Especially here, specific know-how is required to correctly conserve the status of a metabolome of a biological system until it is analyzed.

Spectroscopy, Chapter 7

With the ongoing improvement in resolution and sensitivity, the number of analytes to be assessed simultaneously is steadily increasing. Modern mass spectrometry allows the quantification of hundreds of components from a single sample. Combining MS with suitable

	NMR	FT/IR	Raman	ESI-MS	CE-MS
metabolic fingerprinting cells, body fluids	x	x	x		
metabolic profiling cellular extracts, metabolite fraction	x	x	x	x	x
metabolic target analysis single specific metabolites	x	x	x	x	x

Figure 40.5 General strategies and techniques used to analyze a metabolome.

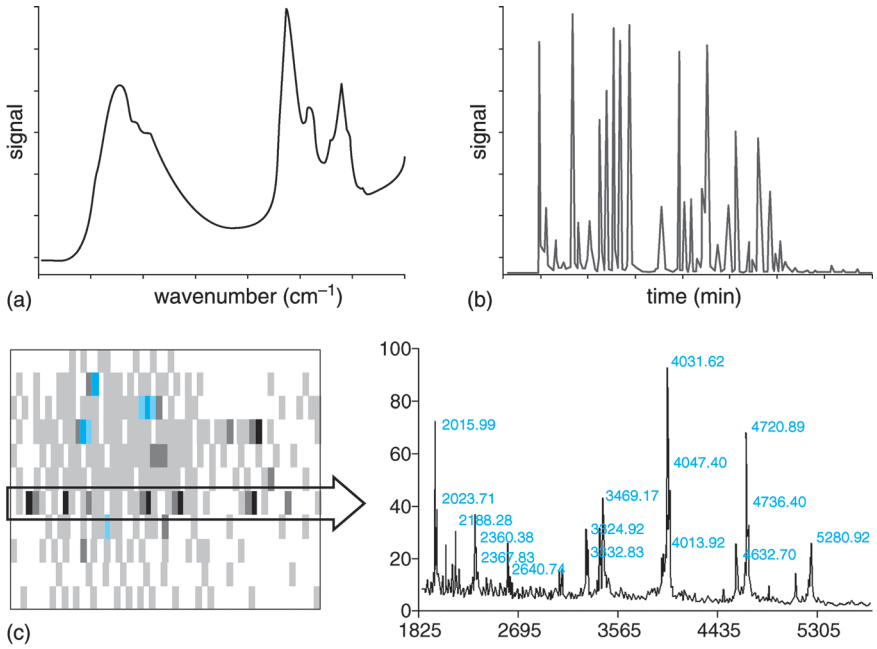


Figure 40.6 Coupling of analytic techniques with separation technologies allows for high separation and resolution power even in complex samples. (a) IR-spectrometry, (b) chromatography, and (c) coupling of chromatography with mass-spectrometry.

separation technologies, thousands of components can be quantified in one single LC/MS experiment (Figure 40.6). A disadvantage of the methods used is the inability to analyze the "*in situ*" state of a high number of metabolites in the living organism. Such "*in situ*" analysis would be seen as the ideal situation for systems biology approaches.

40.4 Peptidomics

The central aim of research interest in *proteomics* is to fully grasp analytical data of proteins from biological samples and turn these data into knowledge. The main technologies used have been evolving from 2D-gel electrophoresis to more protein-digest based methods that are more compatible with high performance mass spectrometry. Several sub-forms of proteomics have evolved dealing with glycoproteins, protein complexes, antibodies, and native peptides. The term *peptidomics* describes the analysis of native peptides in the mass range 1–15 kDa, thus closing the molecular mass gap between metabolomics and proteomics, with a certain degree of overlap with both sides. Peptidomics is today a novel research area in functional genomic analysis, comprehensively analyzing peptides and small proteins at a defined biological state at a defined time point. As already noted, peptides represent true metabolites in the sense that they are products of protein metabolism, usually generated by way of complex protease activities with temporal and spatial resolution. Almost every peptide in an organism is generated by the (specific) cleavage by a peptidase or protease. In the human genome it is assumed that approximately 5% of genes encode for peptidases and proteases, and around 1000 peptidases plus around 200 homologues have been characterized to date (source MEROPS http://merops.sanger.ac.uk). They are located within cells, in body fluids such as blood plasma, urine, and cerebrospinal fluid, and in the extracellular matrix, being able to generate an enormous multitude of peptides.

Cleavage of Proteins, Chapter 9

Peptide species with distinct and relevant biological function have long attracted the attention of a broad spectrum from basic researchers to focusing on medicine. Prime examples are peptide hormones such as insulin from pancreatic endocrine cells or parathyroid hormone from the parathyroid gland. For many other peptides a precise knowledge of their biological function is currently not known, and it is expected that amongst those that are just degradation products ready for clearance or re-uptake a large number of important, bioactive peptides is still to be discovered. For example, collagenous proteins account for 25% of total body protein, and they undergo permanent degradation, remodeling, and alteration processes. They are usually cleaved

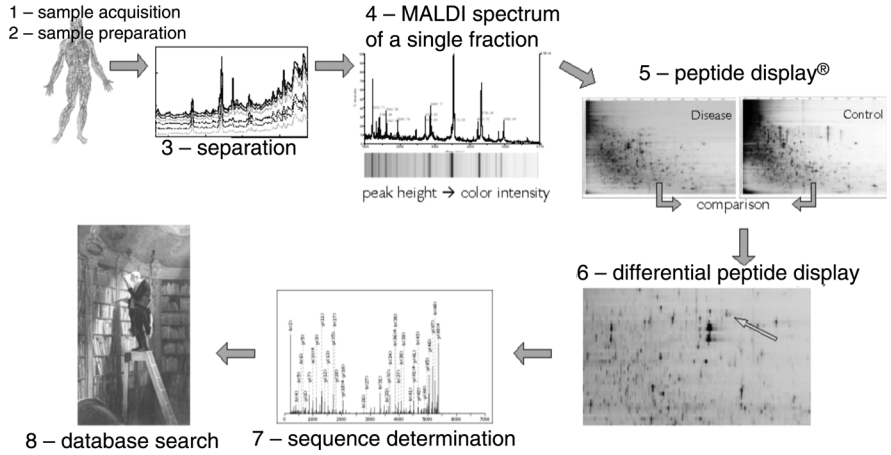


Figure 40.7 Analysis of a peptidome. Following removal of contaminating components (proteins, non-proteinaceous materials such as salts, sugars, lipids, metabolites) the complexity of a peptidome is reduced by sample fractionation techniques. Mass spectrometric analysis is then capable of describing thousands of peptides in a qualitative and quantitative manner. Comparative analysis between sample cohorts facilitates a differential peptide display finally resulting in structural identification of peptide biomarkers.

by collagenases giving rise to molecules such as endostatins, which supposedly alter vascularization of tissues. Of high relevance are intracellular processing events of proteins, shuttling peptides to cell surfaces for immune system priming or converting pre-prohormones into active peptide hormones.

The analysis a peptidome usually requires very detailed, robust, and reproducible methods for qualitative as well as quantitative analysis. Such a process (Figure 40.7) is composed of different, subsequent steps that have to be adapted depending on the biological source and the scientific rationale behind the study. The starting point for a study for peptidomics profiling is the adequate collection and storage of suitable samples. Especially here, the success/failure of subsequent studies is already decided, since inadequate sample pretreatment and storage can never be corrected whatever sound science is applied at later stages.

The second relevant procedure is sample preparation. Biological samples such as body fluids contain sets of analytes with the potential to interfere with the analysis of proteins and especially peptides. During sample preparation one main goal is the removal of such interferences, such as salts, sugars, lipids, metabolites and especially high abundant proteins such as albumins, immunoglobulins, and others. This is achieved via technologies such as size separation by gel chromatography, electrophoresis or ultrafiltration, affinity extraction, and solid phase adsorption. After isolating native peptides they are subsequently fractionated and subjected to mass spectrometric analysis. Detailed analysis of mass spectrometric data, usually in combination with other attributes such as clinical data of a patient, is then performed. To generate knowledge a diverse set of bioinformatical and biostatistical tools is required.

40.5 Metabolomics – *Knowledge Mining*

Knowledge discovery in data is (according to Fayyad *et al.*):

> ". . . the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data".

The collection of large amounts of data is not yet useful information. For this, models have to be created, which can be applied for predictions or alternatively some kind of theory, which can be later used for setting up experiments. A useful model is one that is accurate in the context of the utility of its predictions. A common challenge in omics research is to turn data into knowledge in a systematic way, to integrate previous knowledge and data, and to differentiate between data driven and model driven approaches. Therefore, it is necessary to understand the underlying principles of knowledge and model generation. Profiling technologies, such as metabolomics, usually result in large sets of uninformative data. A key responsibility of a researcher lies in the prioritization and extraction of the correct relevant information and in deciding on a proper experimental design to be able to annotate metabolic profiles. The experimental design already

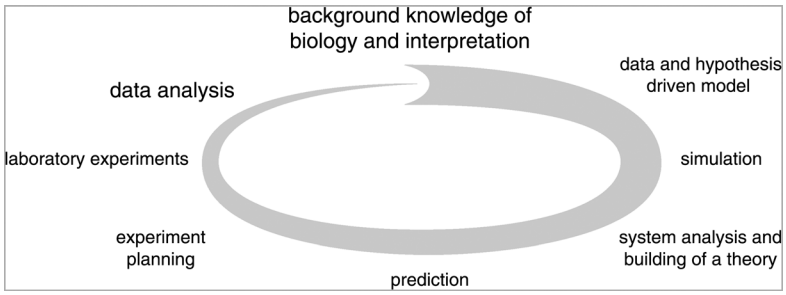


Figure 40.8 Research cycle based on a circular system of hypotheses and knowledge in systems biology. Adapted and modified from Kitano H. (2002) Systems biology: a brief overview. *Science*, **295,** 1662–1664.

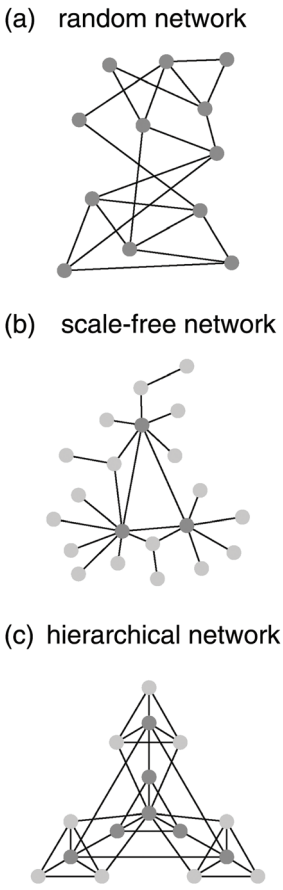


Figure 40.9 Comparison of fundamentally different network architectures. Members in a random network have identical priorities and are therefore connected in a non-hierarchical network. Scale-free networks represent nodes highly connected and therefore superior in relevance to less connected members. They are called scale-free, because zooming in on any part of the distribution does not change its shape. Hierarchical networks demonstrate clear hierarchical organizations. This type of network is usually typical for metabolic pathways in all organisms serving different functional tasks.

contains expectations or prior information. Examples of basic experimental designs are comparisons of data derived from two drastically different states of samples from different individuals (healthy versus diseased) or within an individual (pre- versus post-treatment) to create classifiers. The investigated factors have to be separated from experimental artifacts or nuisance factors, otherwise no relevant and portable knowledge is generated. The metabolomics researcher has to integrate hypothesis driven, data driven, and model driven research and thus create hypotheses based on previous knowledge, extract valid patterns of data, and create and validate models, which have to be compact and can be communicated.

Historically, in classic biochemistry hypothesis-driven deductive work has been the center of activities. With the availability of large profiling data, data driven inductive methods became popular. In purely inductive and data-driven approaches the goal is to model data without reference to pre-formulated hypotheses in order to identify the relevant individual components in an unbiased way. However, the shortcoming of this approach became apparent. One underlying issue comes from the fact that data from fingerprinting and profiling exercises frequently carry measurement errors. To address research questions adequately, creating knowledge, a combination of inductive and hypothesis driven approaches is essential. Existing models are therefore challenged by experiments to test and validate these models (Figure 40.8).

40.6 Data Mining

Analysis of metabolites increasingly delivers high dimensional, multivariate data. In this respect, data of single metabolites, clinical parameters, and properties of analytes represent the different dimensions of an experiment. The data set usually consists of large tables from thousands of analytes in individual sample cohorts. Network analysis intends to reveal a model of correlations in such data, extracting pattern and topology of interaction networks. These networks usually include information on the concentration of whole sets of analytes. The network topology allows us to recognize relevant organizational hierarchy and central molecules as "hubs" (Figure 40.9). One good example of numeric procedures to analyze correlations and associations between metabolites is correlation-associated networks. This procedure utilizes the fact that different metabolites have distinct, very defined relations regarding their concentrations in biological samples. If, now, sets of experiments utilizing different, carefully selected experimental conditions are correlated, the concentration levels of different metabolites can be put into relation towards each other, resulting in mapping of known as well as of novel metabolites towards each other, creating such network topologies.

Data mining is the computational process of systematically handling complex data sets for the extraction of compact patterns or rules. It can be understood as "mining," as a kind of enrichment procedure utilizing methods of statistics, machine learning, and database systems. The common data matrix investigated by data mining methods contains the metabolites and annotation data as attributes and the experimental measurements as observations forming an n-m matrix. The number of formally independent conclusions depends only on the number of observations, which is unfortunately outnumbered by the large number of attributes in "omics" studies. This so-called curse of dimensionality or sparsity of data is problematic for any method that requires statistical significance. Examples of typical methods used for metabolomic data mining procedures are multiple hypothesis testing based on multiple ANOVAs, multiple regression methods, step-wise linear regression, linear mixed effect modeling, multiple logistic

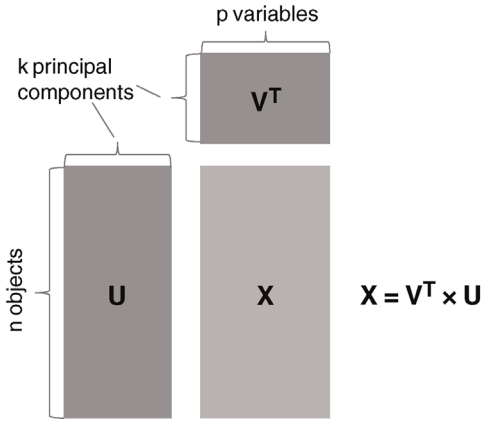


Figure 40.10 Principal component analysis (PCA) subdivides complex data matrices into different projection vectors based on their relative importance (power of component). The product of the most important vector (rows and columns of the first principal component) is utilized to reconstruct the data matrix according to the first component. During PCA the most important components are utilized to denote dominant vectors, leaving behind the influence of the residual vector matrix not represented by the more important vectors.

regression, or the like. More sophisticated methods are derived from machine learning algorithms such as neural networks, support vector machines, Gaussian mixture models, Kohonen network analysis, or other artificial intelligence algorithms. Searching data often relies on detecting areas where objects form groups with similar properties. Conclusions drawn from such underpowered studies are often compromised by random observations. This represents a real challenge for biostatisticians and needs careful consideration in a cycle of experimental replication. In addition, the number of investigated analytes has to be reduced step by step during verification studies to escape this curse of dimensionality.

Powerful methods for data reduction and creating compact models are, for example, the chemometric projection methods for multivariate statistics. They usually aim to extract a few so-called latent variables from the whole measurement set of data, which are shared between all the measured analytes as common metabolic phenomena. A well-known broadly used chemometric method is *principal component analysis* (PCA). Here, the measurement data is transformed in a systematic way to identify so-called principle components, which are related to latent variables (Figure 40.10) and can be interpreted as such.

Latent variables constitute metabolic properties, which usually cannot be observed directly. They represent intrinsic, powerful features important for the understanding of general, relevant sets of parameters of metabolism. PCA thus has the power to identify general factors and biological phenomena and provides a compact model, which can be applied to novel data.

Each new metabolic profile can now be visualized in a coordinate system and the profile is subsequently used for correlation to known metabolic data. Similar metabolic states will group together even if biological specimens are quite diverse (Figure 40.11).

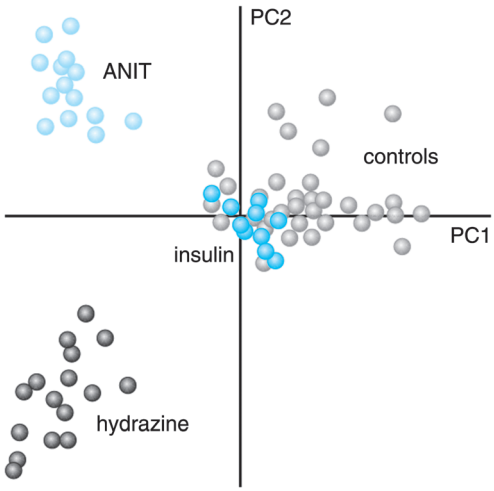


Figure 40.11 Visualization of experimental data of the first two principal components in a simplified coordinate system. Every point represents the measurement of a complex metabolomics profile. Such a simplified coordinate system allows alignment of individual samples along two main, prioritized vectors pointing towards the most important variables in the data space. This can be utilized to interrogate from a top view the degree of self-similarity of different metabolic profiles.

40.7 Fields of Application

The technologies outlined here have shown growing utility in addressing a diverse set of bioanalytical questions:

In biology:

basic research in physiology and pathophysiology of plants, microorganisms, and animals.

In medicine:

- basic research in medicine;
- target identification, verification, and validation;
- measurement of biochemical changes during disease;
- mapping of multiple known as well as novel analytes and correlation to established biochemical pathways;
- generation of novel research hypotheses;
- identification of key enzymes, proteins, and metabolites;
- preclinical studies and toxicological analyses;
- mode of action studies;
- comparison of metabolic profiles with known toxic profiles;
- prediction of toxic effects;
- dose–effect relation mapping;
- extrapolation of species relationships;
- clinical trials;
- classification of subgroups by side effects and adverse events;
- reduction of non-responder populations;
- diagnostics and biomarkers;
- identification of single diagnostic molecules and multiplex panels.

40.8 Outlook

Metabolomics and peptidomics are emerging science fields with a high impact on diagnostics and systems biology. Both fields utilize a set of specific analytic, computerized, and bioinformatics technologies thus merging different aspects of functional genomics. The improved analysis and predictability of living systems reflecting internal and external factors is achieved by metabolomics and peptidomics. Significant impact is expected on progress in medicine, drug development, diagnostics, and basic science. Systems theory and progress in information theory is going to provide a better toolbox for, hopefully, enabling a basic understanding of the miracle of living systems as such.

Further Reading

Beckonert, O., Keun, H.C., Ebbels, T., Bundy, J., Holmes, E., Lindon, J.C., and Nicholson, J.K. (2007) Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat. Protocols*, **2**, 2692–2703.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996) *Advances in Knowledge Discovery and Data Mining*, MIT Press.

Gaenshirt, D., Harms, F., Rohmann, S., and Schulz-Knappe, P. (eds) (2005) *Peptidomics in Drug Development*, Editio Cantor Verlag.

Goodacre, R., Vaidyanathan, S., Dunn, G., Harrigan, G., and Kell D. (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.*, **22** (5), 245–252.

Kell, D. (2004) Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.*, **7**, 296–307.

Nicholson, J. and Wilson, I. (2003) Understanding global systems biology: metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.*, **2**, 668–676.

Roche Applied Science, *Biochemical Pathways*, Elsevier, Heidelberg.

Interactomics – Systematic Protein–Protein Interactions

41

Thomas O. Joos,[1] Stefan Pabst,[2] and Markus F. Templin[1]

[1]NMI Natural and Medical Sciences Institute at the University of Tübingen, Markwiesenstraße 55, 72770 Tübingen, Germany
[2]Pabst, MorphoSys AG, Protein Sciences & CMC, Semmelweisstraße 7, 82152 Planegg, Germany

To gain a more complete understanding of protein-mediated processes in cells, a comprehensive characterization of the proteins involved and identification of their interaction partners is an absolute prerequisite. A major goal of such investigations is not only to find out whether a protein of interest is present, but also to glean information on its functional state. This relates in particular to the protein's localization, potential modifications and interactions with other proteins. The localization of proteins usually involves the application of immunohistological methods and cell fractionation. Protein modifications can be identified using classical approaches such as mass spectrometry and ELISA (enzyme-linked immunosorbent assay). The identification of protein–protein interactions is also of major importance. The central role of such interactions has only been recognized in recent years. However, fundamental biological processes such as DNA replication, transcription, translation, transport, cell cycle control, and signal transduction can only be explained with precise/detailed knowledge of the underlying protein–protein interactions.

Consequently, the study of protein–protein interactions has become a central theme in protein analysis that falls within the field of interactomics. Interactomics deals with all possible interactions between proteins in a given situation. This takes into account permanent multienzyme complexes as well as the association and dissociation of regulatory factors in protein complexes. Protein interaction can be studied using a broad range of different technologies, including two-hybrid interaction analysis, the separation of protein complexes using affinity chromatography, the mass spectrometric analysis of associated factors as well as protein microarrays. The following sections focus on microarrays and present and discuss their potential for use in interactome analysis.

Two-Hybrid System, Section 16.1

Affinity Chromatography, Section 10.4.8

Mass Spectrometry, Chapter 15

41.1 Protein Microarrays

Array systems are analytical systems that enable scientists to perform a large number of measurements simultaneously. An array is an orderly arrangement of microscopic spots, often in rows and columns. The term microarray refers to the miniaturized array format that enables the highly parallel execution of experiments. Protein microarrays can be used to simultaneously identify and quantify a large number of proteins in a single experiment. In addition to studying protein expression, protein microarrays can also be used for global interaction studies and functional analyses in a miniaturized format. The outstanding power of microarray technology comes from the high sensitivity of the measurements along with the ability to analyze tens to hundreds of relevant measurement parameters in a single experiment and in tiny sample quantities.

In a *planar* protein microarray, a collection of capture molecules (e.g., antibodies, antigens, or complex probes such as the lysates of cells or tissues) are immobilized on a solid surface – or

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

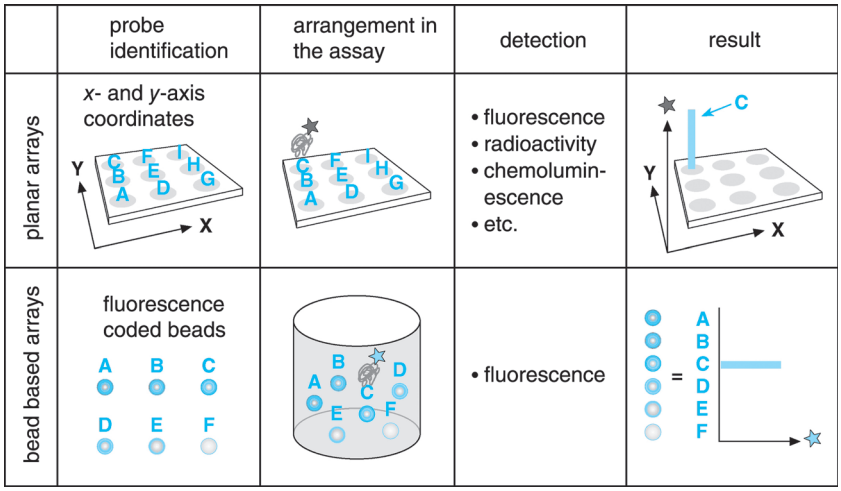


Figure 41.1 Planar and *bead*-based arrays. In a planar microarray, the different probes (shown in blue) can be precisely identified from their *x*- and *y*-axis coordinates. In *bead*-based assays, the distinction of the probes on the beads relies on different *bead* types that can be identified by color (fluorescence) code or size. In planar arrays, the labeled binding partner (gray with asterisk) is incubated directly on the planar chip surface; in *bead*-based arrays, the beads are mixed with labeled binding partners in microtiter plate wells. Different methods are used for the detection of the bound binding partners; *bead*-based arrays usually involve fluorescence-based detection methods.

phase, which is why these microarrays are referred to as solid phase assays. Parallelization is achieved by immobilizing the different capture molecules or probes on planar carriers (e.g., coated glass slides) in so-called microspots (spot diameter: 250 µm) that are arranged in rows and columns, i.e., in arrays. On planar surfaces, the identity of the immobilized capture molecules can be determined from their *xy* coordinates (Figure 41.1).

In analogy to the orderly spatial separation of planar microarrays, the ability to analyze different immunoassays in bead-based systems requires the use of other codes for their unique identification. For example, miniaturized and parallelized ligand binding assays employ color-coded or size-coded microspheres where the probes are attached to spectroscopically distinguishable fluorescence-coded beads only a few micrometers in diameter (e.g., polystyrene *beads*). Arrays are created with microsphere beads of different sizes or colors to which different probes are attached (suspension arrays). The molecules that are attached to the beads can be identified from the microspheres' unique identity based on variations in fluorescence (Figure 41.1). The different color codes can be effectively identified using a technology that enables the separation of the beads in the suspension on the basis of their optical properties (FACS, fluorescence activated cell sorting). This technique was originally used for cell analysis. Cells in suspension or, in the case of suspension arrays, color- or size-coded microspheres in solution are aligned in a stream of fluid so that they pass single file through the detection cell. Two lasers of different wavelengths classify the different beads and determine the quantity of ligand bound. The bound ligands are usually identified using fluorescence-based methods.

FACS, Section 5.5

41.1.1 Sensitivity Increase through Miniaturization – Ambient Analyte Assay

The array systems described here have two principal characteristics in common: the high degree of parallelization and the associated extreme reduction of the area available for analyte detection. This miniaturization was the major priority of the first experiments done in this field that nowadays are referred to as microarrays. In the 1980s, Roger Ekins was looking for conditions under which immunoassays/ligand binding assays would achieve the best possible sensitivity. His ambient analyte assay theory outlines the consequences of miniaturization.

Capture molecules (e.g., antibodies) are immobilized on a small area – the microspot – on a solid phase. Although the total number of capture molecules (antibodies) per microspot is relatively low, they are nevertheless arranged in a very high density. These capture molecules form complexes with their target molecules. The number of target molecules bound is relatively low as it is identical to the number of capture molecules contained in a microspot. The number of free target molecules in the sample therefore does not change. This is also the case with a low/small number of target proteins. Assays in which a microspot contains $0.1/K$ or less of capture molecules are referred to as ambient analyte assays (K is the affinity constant of the binding reaction). The proportion of bound target protein correlates directly with the concentration of

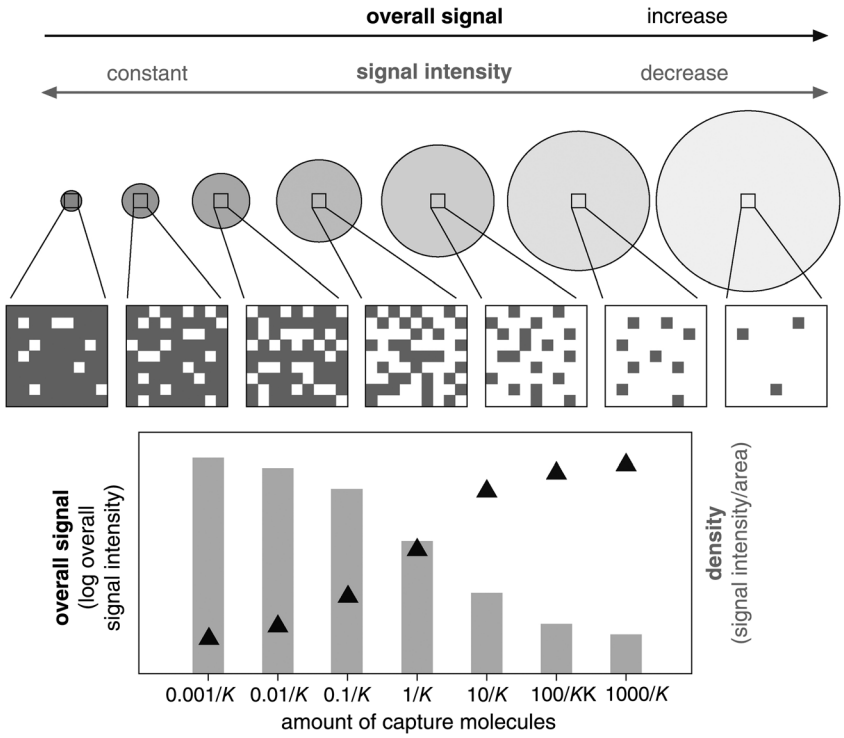


Figure 41.2 Sensitivity increase through miniaturization – signal density and overall signal in microspots (adapted and modified from Ekins, R. and Chu, F. (1992) *Ann. Biol. Clin.*, 50, 337–353). The course of signal density (signal intensity/area, grey bars) and overall signal (signal intensity, black triangles) was determined for microspots with increasing quantities of capture molecules. As all microspots have the same capture molecule density, larger microspots contain larger quantities of capture molecules. The overall signal (overall signal intensity) therefore increases with increasing spot size and reaches a maximum when all target molecules contained in the sample are bound in the microspot. For microspots with fewer capture molecules and smaller spot size, the signal density (signal intensity/area) increases and reaches an almost constant value when the capture molecule concentration value is below 0.1/*K* (*K* = affinity constant). These conditions characterize ambient analyte assays. The concentration of free target molecules in the sample barely changes despite the fact that they form complexes with the capture molecules that are immobilized in the microspot. In the figure, signal intensity and the derived area ratio (signal density) are depicted.

target protein in the sample. In addition, under such conditions, the results are insensitive to the volume of sample used. Ekins' ambient analyte assay theory therefore relates to the increased sensitivity of miniaturized immunoassays.

This increased sensitivity is due to two phenomena:

- the reaction of complex formation occurs at maximum target molecule concentration;
- capture–target molecule complexes are formed on an extremely small area – the microspot – which leads to a high local signal intensity (Figure 41.2).

The following illustrates this relationship. On a surface, microspots of increasing size and constant concentration are produced from a capture molecule solution. Therefore, the total number of immobilized capture molecules increases the greater the spot area becomes. As a consequence, the overall signal of the respective microspots also increases. However, as the target molecule is not present in unlimited quantities in the sample, the signal density will decrease as the spot area increases. The formation of complexes consisting of capture molecule and target molecule leads to a reduction in the quantity of free target proteins in the sample at the same time as the complexes formed are distributed across a larger area. A larger spot therefore exhibits a lower maximum signal. Although lower overall signals will be observed for smaller microspots, the signal density (signal intensity per area) increases (Figure 41.2). The signal density will reach an optimum value below a certain spot size and will not increase further. Under ambient analyte assay conditions, the quantity of target protein does not represent a limiting factor.

41.1.2 From DNA to Protein Microarrays

The development of DNA chip technologies has made it possible to analyze large numbers of probes in a single experiment, which is achieved by immobilizing a large number of different capture molecules on a carrier. In addition, the search for conditions that enabled immunoassays to exhibit maximum sensitivity led to the miniaturization of the system. Ekins and his colleagues showed that the use of miniaturized systems enables the sensitive detection of TSH (thyroid stimulating hormone) or HbsAg (hepatitis B surface antigen) including in

Table 41.1 Comparison of DNA and proteins in terms of their microarray suitability.

Characteristics	DNA	Protein
Structure	Uniform Hydrophilic Stable	Varies Hydrophilic and/or hydrophobic domains Very sensitive to stable
Functional state	Denatured, no activity loss → can be stored dry	3D structure is crucial for activity, denaturation should be prevented → needs to be kept moist (in a buffer)
Interaction sites	1 : 1 Interaction	Several interaction sites
Interaction affinity	High	Varies from protein to protein; very low to high
Interaction specificity	High	Varies from protein to protein; very low to high
Prediction of activity	Easily possible, based on nucleotide sequence	Not (yet) possible, but bioinformatic methods are being developed (based on sequence and structural homologies)
Amplification method	Established (PCR)	Not (yet) available

femtomolar concentrations (10^6 molecules per ml). These investigations showed that miniaturized systems have huge potential for application in basic research and medical diagnostics. As DNA microarray technologies have long been established as important and well-functioning laboratory methodologies, many attempts have in recent years been made to transfer this technology to protein microarrays. It has been established that the transfer of the technology from DNA to proteins is theoretically possible, but solutions still need to be found for numerous issues. The associated challenges are more of an intrinsic than of a methodological nature, and arise from the particular requirements of proteins.

As far as the equipment used to produce DNA and protein microarrays is concerned, the same devices or devices with just minor technical modifications can be used for printing planar protein microarrays on a microscope slide. Different systems, including needle-based contact printers (split-needle or pin-and-ring systems) and contactless microdispensing systems (inkjet systems), can be used to automatically deposit nano- or picoliter sample volumes in columns and rows on suitable carrier materials. Protein interactions can be detected using confocal laser scanners such as those used for the analysis of DNA microarrays.

While the challenges associated with the generation of protein arrays are less technical, the properties of proteins make global approaches involving thousands of capture molecules appear far more difficult than for DNA chips. DNA molecules are a homogeneous class of molecules with fairly similar physicochemical properties due to the DNA's defined sequence of four nucleotides joined together by an acidic sugar backbone. In contrast to DNA molecules, proteins are much more complex (Table 41.1). They consist of 20 different amino acid building blocks that can have different conformations (secondary, tertiary, and quaternary structures), different charges (no charge, positive, or negative net charge), and solubilities (hydrophilic or hydrophobic). In addition, proteins are far less robust than DNA molecules and have a low level of stability. It is therefore important to immobilize the proteins on a substrate without compromising their function, which is an important prerequisite for analyzing protein–protein interactions. In contrast to DNA–DNA interactions, which, due to the complementary pairing of the nucleotides, can be accurately determined, it is currently impossible to predict from amino acid information alone the amino acid sequence a potential interaction partner needs to have in order to be able to bind specifically to the immobilized capture molecule. In addition, the generation of the required capture molecules is frequently a time-consuming and laborious process; no PCR equivalent is available that would allow the amplification of proteins. And last but not least, the interactions of proteins depend on a large number of buffer conditions, including the pH value, salt concentration, and cofactors required. It is therefore quite difficult to develop a universal protein array system in analogy to DNA chips that would take into account the individual requirements of the proteins and would make it possible to analyze as many interactions as possible under physiological conditions. Care must therefore be taken to interpret the protein microarray results while keeping in mind the particular experimental methodology used.

41.1.3 Application of Protein Microarrays

Protein microarrays can provide qualitative as well as quantitative information. They provide information on whether a protein is present in a biological sample. For example, tissue samples of different origin (tumor tissue and normal tissue) can be investigated to find out in which of the two samples a certain protein is present or absent. However, in most cases, quantitative information is also required. Information of this type might relate to the difference in the amount of marker protein in the two samples. In principle, protein microarrays can be used for identifying any molecular interaction in which two partner molecules specifically recognize one another. In addition to protein–protein interactions, these are also antigen–antibody reactions, enzyme–substrate, and ligand–receptor interactions. Microarrays can be produced from complex samples such as cell and tissue lysates as well as from whole and intact cells. The following subsection deals with different protein microarray technologies that highlight the enormous versatility of miniaturized assay systems (Figure 41.3).

Protein–Protein Interaction Protein microarrays are excellent tools for the parallel analysis of protein–protein interactions. Firstly, they are used to identify new interaction partners of a specific protein (screening tool). Numerous protein microarrays for analyzing protein–protein interactions are commercially available. One of these is a planar array developed by Invitrogen, which is a standard size nitrocellulose-coated glass slide on which more than 12 000 protein spots around 150 µm in diameter, more than 4000 different proteins of the yeast *Saccharomyces cerevisiae*, and more than 2000 control proteins are immobilized in duplicate. A single experiment with this array therefore enables the identification of potential interaction partners for almost the entire yeast proteome (approximately 6500 proteins).

An example of an interaction partner identified with the Protoarray protein microarray (ThermoFischer Scientific) is shown in Figure 41.4. The yeast protein MOG1, which is involved in nuclear transport, was used as probe. MOG1 was biotinylated prior to use in order to enable the subsequent detection of the reaction with fluorophore-conjugated streptavidin. MOG1 interacted well with the protein GSP1, a GTP-binding protein (yeast homologue of mammalian Ran) that is also involved in nuclear transport.

Antibody–Antigen Interaction Protein microarrays are often used for analyzing antigen–antibody interactions. An antibody microarray usually contains a broad range of antibodies with different specificities. The chip is then probed with a mixture of proteins that are labeled with a dye (e.g., fluorophore-labeled tissue lysate). This enables a large number of antibodies to be tested simultaneously with a very limited amount of sample.

Immunological Techniques, Chapter 5

Miniaturized sandwich immunoassays have proven to be especially suited to the sensitive analysis of analytes. Using the same method as a classical ELISA, capture molecules are spotted and fixed on a solid surface. The analyte in a liquid sample (liquid phase) is bound by the capture molecules and detected by a second antibody that also has an analyte binding site. Miniaturization has the advantage that several analytes can be analyzed and quantified simultaneously with the same sensitivity as classical ELISAs.

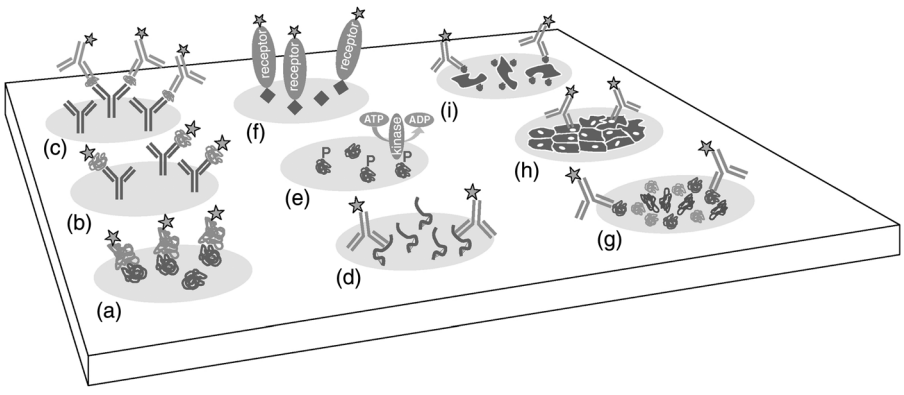


Figure 41.3 Schematic representation of different protein microarray applications. (a) Set-up used for the identification of protein–protein interactions. A protein (dark grey) that is immobilized in the microspot reacts specifically with another protein (light grey) which is labeled with a fluorophore (asterisk) for detection. (b) An antibody array in which the antibody is immobilized on the chip surface and captures a labeled analyte from the solution. (c) A sandwich immunoassay, where the analyte is captured by an immobilized antibody, and detected with a second antibody that has another binding site on the analyte. (d) Illustration of a peptide array such as those used for the characterization of peptide-specific antibodies. Protein microarrays can also be used for the detection of enzyme–substrate interactions (e) and ligand–receptor interactions (f). Reverse-phase protein microarrays can be produced with cell and tissue lysates (g), tissue sections (h), or whole cells (i).

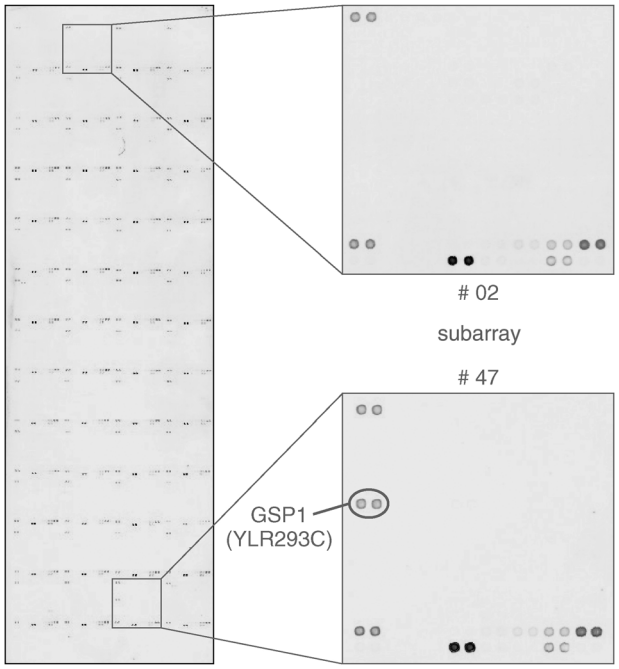


Figure 41.4 Planar Protoarray microarray, which contains more than 4000 different yeast proteins. The proteins on the microarray are printed in 48 subarrays that are equally spaced vertically and horizontally (dimensions: four columns × twelve rows). Each subarray consists of 256 microspots, arranged in 16 columns and 16 rows. The entire array contains more than 12 000 protein microspots. The protein array was incubated with the yeast protein MOG1, which plays a crucial role in nuclear transport. The protein was biotinylated prior to use, enabling bound MOG1 to be detected using fluorophore-conjugated streptavidin. The left-hand side of the figure shows the entire slide with all subarrays; the right-hand side shows two enlarged subarrays (# 02 and # 47, with a dark grey frame). Signals that occur in both subarrays are controls. The signals in the dark grey circle depict the interaction of MOG1 with GSP1 (YLR293C), a nuclear import protein that was immobilized on the chip surface.

Instead of antibodies, it is also possible to spot and fix antigens on a solid surface and identify them using a labeled antibody. Whole protein molecules, short protein fragments or peptides can be used as antigens. Peptide arrays can be used for the identification of antibody epitopes (antigenic determinants), i.e., the residues of an antigen that are crucial for antibody binding. The antigen sequence needs to be known for this purpose. The process starts with the chemical synthesis of peptides (10–20 amino acids long) that represent the entire antigen sequence. Ideally, peptides with partially overlapping sequences are generated. The peptides are immobilized on a suitable surface and the peptides that interact with an antibody of interest are identified. Another prerequisite for this method is that the antibody needs a high enough affinity for it to bind to its binding site. The method is therefore better suited to the mapping of continuous (linear) epitopes than to the identification of discontinuous epitopes (conformational epitopes).

Enzyme–Substrate Interaction Enzyme–substrate interactions are studied using enzyme substrates that are spotted and fixed in arrays on suitable surfaces. The array is then incubated with the enzyme under investigation. Such microarrays can be used to analyze different substrates for their protein kinase specificity (substrate profiling). Substrate phosphorylation is usually detected using a phospho-specific antibody that only binds phosphorylated substrates or a radioactively labeled ATP and a phosphoimager.

Enzyme Activity Testing, Chapter 3

Ligand–Receptor Interaction To study ligand–receptor interactions, membrane fractions or low-molecular organic substances containing receptors are immobilized on a microarray and incubated with labeled binding molecules. The binding behavior of substances to receptors can thus be analyzed simultaneously under identical conditions. Such array systems are of major interest for the pharmaceutical industry in its search for new drugs (drug screening), as miniaturization and parallelization minimizes the consumption of reagents at the same time as enabling an increase in the number of parameters that can be studied per experiment.

Reverse Phase Microarrays Reverse phase protein microarrays contain protein fractions that are generated by cell lysis or tissue microsection. Each microspot either represents the proteome of a tissue sample that correlates with different disease stages or the proteome of healthy tissue.

Suitable antibodies are used to study these proteomes for the presence of molecular particularities associated with a certain disease. The results of such investigations might be

of particular benefit in the future as they may make it possible to tailor patient therapy to individual requirements.

Microarrays can also be produced from different tissues. Tissue arrays allow the rapid and effective investigation of protein expression patterns with suitable binding molecules. In addition, they enable the effective profiling of antibodies for their suitability for application in histological investigations or as disease markers.

In addition to lysates or tissue sections, it is also possible to immobilize whole cells on solid carriers. Such cell arrangements (cell arrays) are excellently suited to the characterization of antibodies that are directed against cell surface molecules (e.g., MHC molecules) and for the comparative analysis of cell surface molecules of different cell lines.

Further Reading

Ekins, R.E. (1989) Multi-analyte immunoassay. *J. Pharm. Biomed. Anal.*, **7**, 155–168.

Ekins, R. and Chu, F. (1992) Multianalyte microspot immunoassay. The microanalytical 'compact disk' of the future. *Ann. Biol. Clin.*, **50**, 337–353.

Templin, M.F., Stoll, D., Schwenk, J.M., Pötz, O., Kramer, S., and Joos, T.O. (2003) Protein microarrays: promising tools for proteomic research. *Proteomics*, **3**, 2155–2166.

Chemical Biology

42

Daniel Rauh and Matthias Rabiller

Technische Universität Dortmund, Chemische Biologie, Otto-Hahn-Straße 6, 44227 Dortmund, Germany

42.1 Chemical Biology – Innovative Chemical Approaches to Study Biological Phenomena

Sequencing the human genome as well as the genomes of a growing number of other organisms and the development of powerful "-omics" technologies have provided a new basis for the study of cellular processes. Science is now confronted with the daunting task of translating these technological breakthroughs and the resulting genomic and proteomic data into useful knowledge. In this context, considerable hope is placed in biology to develop new approaches to address the challenges of the twenty-first century: energy and food supply, innovative materials, and drug development. The first successes have already signaled biology's potential: crops can be genetically modified to preserve yields despite declining water supplies, microorganisms can be genetically modified to produce large amounts of biofuel or hydrogen. The treatment of human diseases such as cancer has also experienced significant successes based on the decoding of genetic information, which allows new therapeutic approaches, such as personalized medicine. For example, about 10% of the patients suffering from non-small cell lung cancer (NSCLC) carry a mutation in which leucine 858 in the cellular kinase domain of the epidermal growth factor receptor (EGFR) is replaced by an arginine residue, leading the mutated EGFR to continuously promote cell division. Cellular mechanisms that should inhibit uncontrolled growth fail: a tumor starts growing. With the understanding of the mutation as the cause of the tumor and that the cells that harbor this mutation are highly dependent on this receptor signal for their continued growth, a process referred to as oncogene addiction, came the possibility for drugs that specifically block the function of the mutated receptor, once sequencing has established its presence in an affected patient.

Even though many of these promising approaches are still far from general use, they pave the way for future developments. What these approaches all have in common is, however, that the function of the native or mutated gene products, the proteins, is understood in detail at the cellular level. Only then it is possible to analyze and utilize the connections between the sequenced genotype and the observed phenotype. The rapid developments in the last few years in the field of life sciences have led to the study of more and more biological phenomena on the molecular and cellular levels. The challenge inherent in this task is apparent. The key is the realization that many questions can only be addressed successfully at the interface between chemistry and biology with the aid of their often different, but fundamentally complementary, approaches and methods.

The goal of chemical biology is to solve these fundamental biological problems through the development and application of innovative methods from the broad array of techniques available in chemistry (Figure 42.1). Often, chemical biology approaches have their roots in the analysis of structural biological data, which are the manifestation of biological

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

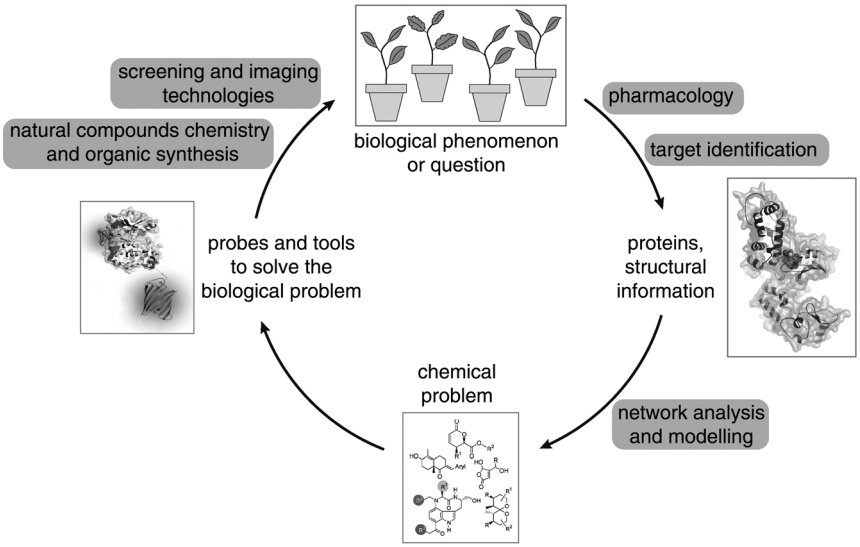


Figure 42.1 Chemical biology is a cyclic process in which biological phenomena are addressed. Most biological phenomena arise from orchestrated protein functions that regulate these processes taking place inside the cell. Chemical biology uses custom made molecules to perturb these interactions (often rooted in molecular networks) specifically and analyzes their role. The development of organic molecules used in biological experiments faces their designers with the challenge of identifying chemical processes that yield the desired molecular properties while respecting the highly sensitive intracellular environment and/or its isolated components.

phenomena. From the fundamentals of biochemistry and molecular cell biology, we know that proteins and their dynamic regulation are central elements of living cells. As a result, biological phenomena can often be traced back to the proteins that are responsible for them. Against this background, proteins, as well as their post-translational modifications, are the particular focus of chemical biology, as well as sugars and lipids.

Protein–protein interactions or the interaction of a given protein with a particular effector or inhibitor molecule can serve as a useful starting point for chemical biology research. Based on this information, new approaches have been developed with which chemical biology is used to study the biological phenomenon, usually by the use of preparative organic chemistry to generate the suitable reagents or molecular probes. An almost infinite repertoire of molecular probes is conceivable. Such probes can be used as reporter groups and as tags to mark proteins in cells, or as small biologically active molecules that can serve to modulate protein function specifically. The probe molecules developed are characterized by means of biochemistry and biophysics before they are used in mostly cell biological experiments. The knowledge gained by the application of these tools not only actively supports further development cycles within the chosen chemical biological method and contributes to a better understanding of the biological phenomena defined at the outset, it also includes the potential to be useful as starting points for modern drug development.

It becomes clear that chemical biology is a multidisciplinary approach that combines methods from related sciences such as chemistry, biology, and physics. Practically, these interactions take place through collaboration, either within a laboratory or between complementary laboratory groups. The spectrum of methods used in chemical biology covers a very broad range and is becoming ever larger. A few of the key aspects, some of which will be addressed later in this chapter, are listed briefly here:

- Methods for the identification and organic synthesis of biologically active molecules that are suitable for use as modulators and molecular probes for perturbing biological systems.
- Methods for the design and the development of semi-synthetic proteins and reporter molecules, which can be switched on or off by the use of drug-like substances or external light.

Fluorescence Spectroscopy, Section 7.4, 32.5

- Methods for the development of meaningful biochemical and cellular assay systems to allow the screening of substance libraries. In particular phenotypic assays and high-content screening are in focus. Use of modern fluorescence techniques like time-resolved fluorescence (FRET/FLIM) and fluorescence correlation spectroscopy (FCS) allow the visualization and analysis of time- and spatial-dependent effects in living cells.

FRET, Section 16.7, 32.5.3

Stable Isotope Labeling in Proteomics (IACT, SILAC), Section, 39.6

- Identification of the targets of a given biologically active molecule. In particular, powerful mass spectroscopy (ICAT and SILAC), as well as phage- and ribosome-display, are methods of choice.

- Methods or techniques for the validation of previously identified proteins in the full breadth of their biology.
- Powerful methods and techniques of chemo- and bioinformatics, for example, to analyze the correlation between a given chemotype and the observed phenotype.

In addition to classic organic chemistry, structural biology, and mass spectroscopy, methods from nano- and micro-technology are increasingly coming into use. As a result of the wide scope of materials and methods, this chapter can only offer an introduction to the broad field of chemical biology, and its utility is illustrated by selected examples. The reader is referred to the Further Reading at the end of the chapter for additional information.

42.2 Chemical Genetics – Small Organic Molecules for the Modulation of Protein Function

Over the last two decades, numerous methods for the investigation of biological phenomena on the molecular and cellular levels have been developed. They are based on the targeted manipulation of the structure, and therefore activity, of proteins. The discovery of the pleiotropic functions of proteins, particularly within complex molecular networks, was made possible through the use of genetic methods such as mutagenesis, gene knock-out and knock-in experiments, and molecular reporters such as the green fluorescent protein (GFP).

GFP, Section 7.3.4

This post-genomic era has also found that classic molecular biological approaches are often not able to close the gap between a measured genotype and an observed phenotype. This is because proteins in cells are rarely independent functional units. Instead, proteins usually manifest their functions as parts of complex structures like cascades and networks, which often interact with and regulate one another. In addition, identical proteins can carry out different functions, dependent on its context in different protein networks. In such networks, a large number of proteins interact with one another physically and chemically. The complex temporal and spatial composition of such systems, as well as their dynamic regulation, represent the individual characteristics of a biological process, such as cell division and differentiation, and are essential for life. It is, therefore, apparent that the quantification of the information, which flows through such networks, is one of the great challenges in modern cell biology. Hopefully, such analyzes will not only provide a better picture of which proteins interact with one another but also reveal how their dynamic regulation is resolved in time and how the corresponding functions emerge from them. Clearly the purely genomic or proteomic information of biological processes is insufficient to describe the complete function of a protein in the context of a cellular network of a living cell, a complete organism, or even just a disease process.

One of the most powerful approaches to the decoding of complex biological questions is the targeted perturbation of protein functions and the differential analysis of perturbed and unperturbed states (Figure 42.2). This is comparable to the approach of an engineer, who first takes an unknown machine apart and then puts it back together, leaving out individual parts, in order to understand the function of the apparatus and its components. In modern cell biology, functional perturbations are achieved through techniques such as mutagenesis, gene knock-out/-in, or through siRNA. These methods have often proven their utility and allowed the analysis of entire genomes in modern high-throughput approaches. Techniques to control

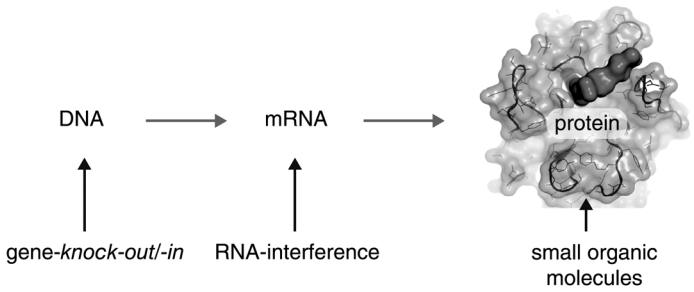


Figure 42.2 Study of the function of proteins. The functions of proteins can be investigated on various levels with the help of perturbations. All the methods shown have their advantages and disadvantages. While gene knock-out and RNA interference physically remove the protein from the cell by controlling its translation or transcription, small organic molecules allow the study of the target protein in its native physiological environment.

Table 42.1 Use of biologically active molecules for the modulation of protein functions has decisive advantages relative to other methods.

Parameter	Advantage
Fast action/temporal control	Small biologically active molecules work rapidly and can be used at any time point in the development of a cell or organism
Spatial control	Through the use of suitable physical (e.g., microfluidics systems) or chemical modifications (e.g., affinity to membranes, pro-drug approaches) the effects can be temporally and spatially confined to defined compartments
Reversibility	Substances are subject to pharmacological processes in a biological system, such as diffusion, metabolism, and excretion. The biological effect of a small biologically active molecule is therefore usually reversible
Dosing	By testing different concentrations of the small molecule, the graded phenotypes can be controlled and quantified
Differentiation	Substances allow the differentiation of the function of protein variants that come from the same gene
Minimally invasive	Substances allow studies on native systems and, therefore, are considered minimally invasive

proteins on the transcriptional level are, however, not trivial and, particularly when animal studies are the focus, are still a great challenge. The development of a transgenic mouse or a knock-out mouse takes months, and the use of RNA interference and the targeting of cells by siRNA requires optimization steps for each sequence used, and every cell type or organism. In addition, RNA interference eliminates the synthesis of proteins but does not influence the pool of the target protein that has already been synthesized in the cell. The targeted "switching off" of a protein function (e.g., of a given enzyme) depends on its natural turnover rate and, in the most unfortunate cases, can last hours or days. The resulting temporal delay between the induced genotype and the observed phenotype allows the cell time to adapt and to compensate for the provoked changes, for example, through the activation of alternate network structures. A further fundamental problem of RNA interference and knock-down techniques is that the changes are chronic, and the protein under investigation is permanently physically removed from the cell. Particularly in the case of enzymes, not only is the originally targeted enzyme activity lost, other important secondary functions, such as scaffolding and the participation in protein–protein interactions, are also lost. Many proteins only manifest their functions through the formation of complex structures with other proteins. Permanent removal of one of these components from the cell can endanger the integrity of an entire protein complex. In addition, a detailed study of protein function over a defined time frame or at a particular time point in the development of an organism is difficult to achieve. Analogous problems result when the gene product under investigation plays a critical role in the developmental process of the cell or the organism, and the modification of the gene is lethal in the embryonic phase.

RNA Interference, Section 38.3

Modern cell biology, therefore, needs methods that allow for the perturbation of protein functions in a biological system, without disturbing the spatial and temporal distribution of the protein of interest. As we will see, the use of small organic molecules as probes for the targeted perturbation of protein function is an attractive and often complementary approach to the genetic methods (Table 42.1). In addition, chemical and genetic approaches can be combined to improve selectivity in targeting a particular protein function.

42.2.1 Study of Protein Functions with Small Organic Molecules

From a pharmacological and chemical point of view, drug-like small molecules are ideal tools for the accurate perturbation of a target protein or of its interactions with other partner molecules inside a living cell. For example, enzyme inhibitors allow interference with the activity of the enzyme at any chosen point time in the development of a cell or an organism. With the aid of chemical perturbation of biological systems, not only can the connections in a protein network be discovered, it also allows the study of highly dynamic processes like the cytoskeleton reorganization, where classical genetic approaches no longer function due to the inherently

dynamic nature of the system. The rapid onset of the pharmacological effect of a biologically active compound also helps to get around the lack of an observable phenotype, which is occasionally observed under the control of single gene knock-outs due to time-dependent transcriptional compensation. In addition, the biological effect of a substance is frequently reversible, and the use of different concentrations allows for the induction of grated phenotypes, which together allows for the transient temporal control of a system. This perspective shows the necessity of the development of universal methods with which the function of virtually every gene in the genome can be modulated by the use of small organic molecules. Ideally these methods should be complementary to genetic approaches. The systematic use of biologically active molecules for the study of protein functions is, therefore, a central discipline of chemical biology.

Chemical biology and the perturbation of complex systems with small molecules can be traced back to the use of poisonous natural substances for the study of biological processes. A few of these molecular probes, like brefeldin (study of vesicular transport in cells), okadaic acid (phosphatase inhibitor), or the alkaloid colchicine from the autumn crocus (study of cell division) had a fundamental impact on entire fields within biology. The discovery of these natural substances established a new paradigm involving the use of molecular probes. Advances in the fields of organic chemistry (e.g., automated parallel and solid-phase synthesis and new approaches for the total synthesis of complex natural substances), screening technologies (automation and miniaturization of the screening of substance libraries, powerful visualization technologies, as well as the development of meaningful phenotypic assays), protein technologies (design of conditional systems for the activation and inactivation of mutants), and computer sciences (chemo-informatics, bio-informatics, the use and management of large data sets such as those from screening campaigns) have expanded the paradigm to allow a systematic approach to the search, design, and development of drug-like molecules that are suitable for the study of biological systems.

The success of chemical genetic approaches is dependent on the discovery and availability of organic molecules, which bind to a protein or protein complex as a ligand and thereby modulate their function. Suitable molecules are usually found during the screening of large substance libraries that are designed and synthesized for, for example, biological relevance (focused libraries based on biologically active precursors), similarity to drugs, structural chemical diversity (to cover as much chemical space as possible), or targeted based on structural approaches. Table 42.2 provides an overview of the approaches used to create substance libraries. The substance libraries are tested in several biochemical and/or phenotypic assay systems for their biological activity. The two fundamental approaches are forward chemical genetics and reverse chemical genetics (Figure 42.3). In the latter approach substance libraries

Table 42.2 Source of substance libraries. Besides several different possibilities for the design and synthesis of substance libraries, the following three complementary approaches have proven to be particularly useful for chemical biology research.

Approach	Use
Biology-oriented synthesis (BIOS)	The design of substance libraries based on natural substances. In the course of evolution, Nature has created a series of structurally complex and potent natural substances that control the function of proteins and downstream biology. Adrenalin and estrogen are good examples, which carry out their functions as biological messengers that interact with cognate receptors. But also plant components like alkaloids, which play an important role in the defense against pests, rate as good starting points for the synthesis of substance libraries
Diversity-oriented synthesis (DOS)	This approach focuses design and synthesis on covering as much chemical space as possible
Structure-based ligand design (SBLD)	This approach takes advantage of structural information (e.g., 3D structural information from protein X-ray crystallography) available about the target protein to guide the design and synthesis, which allows the rational development of small molecule modulators (e.g., inhibitors) and offers starting points for orthogonal, chemical-genetic approaches like the bump-and-hole approach (Figure 42.6)

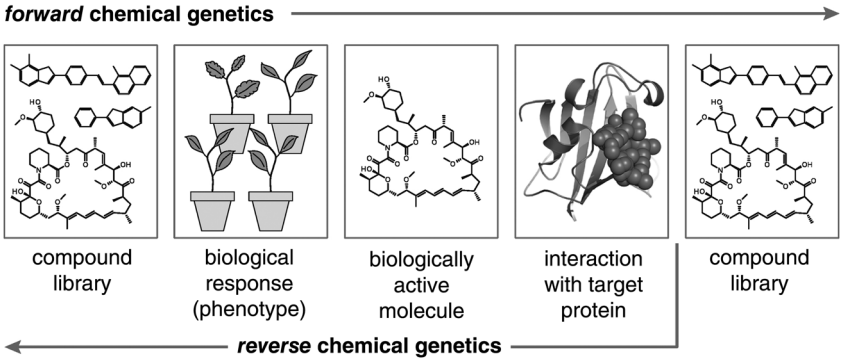


Figure 42.3 Chemical genetics. *Forward* (to find the protein responsible for a particular function) and *reverse* (to find the functions of a particular protein) chemical genetics approaches start by screening substance libraries for biologically active compounds. Substance libraries can be acquired by various means. Besides preparative organic synthesis and the isolation of natural substances from plants, microorganisms, or animals, today libraries of millions of compounds can be purchased.

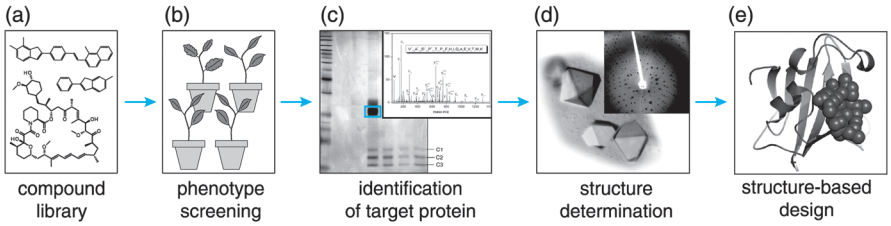


Figure 42.4 Chemical genetics in the case of the protein–protein interaction stabilizer fusicoccin. (a) The influence of individual members of a substance library on the growth of plants is investigated. (b) A few of the tested substances show an obvious phenotype. (c) The biologically active molecule discovered this way was chemically modified and attached to a suitable carrier to enable the successful identification of its target protein by means of two-dimensional electrophoreses and mass spectrometry. (d) Structural biology enabled detailed understanding of atomic interactions between biologically active compounds and the identified target protein. (e) In this example, the X-ray structure enabled identification of the active substance as a stabilizer of protein–protein interactions. The discovered complex natural substance (transparent space-filling model) stabilizes the protein–protein interaction consisting of an adapter protein (gray surface) and the target protein (blue space-filling model). The complex structure of the ligand and protein served as a starting point for the rational design of new small molecules with optimized characteristics (affinity, selectivity, etc.). Through the coordinated efforts of chemistry and biology, new compounds can be identified and optimized which, due to their wilting activity, for example, can be further developed to combat weed growth.

are usually screened against to a predefined target structure (e.g., the inhibition of enzyme activity or protein–protein interactions) and the resulting substances were then used as probes for the investigation of the function of the target protein in physiological systems.

In forward chemical genetics, the target structures are unknown at the outset. Substances responsible for desired effects or a new interesting phenotype found while screening in cell-based assays are later used for the identification of target proteins (e.g., by means of pull down and mass spectroscopy). Figure 42.4 outlines this approach with a stabilizer of a protein–protein interaction, namely, fusicoccin.

42.2.2 Forward and Reverse Chemical Genetics

Genetic analysis can be divided into forward approaches, in which certain phenotypes are analyzed and reverse approaches, in which the phenotypic consequences of mutations are in focus. Forward approaches aim to identify the modulated gene product, which is the cause of the phenotype. Analogously, reverse chemical genetics is based on phenotypes induced by small organic molecules, which interact with certain biological macromolecules like proteins, RNA, or DNA. Reverse chemical genetics often starts with a known protein (it may have been identified by genetics, molecular biology, or earlier chemical biology experiments) and seeks to investigate its function using perturbation experiments. To this purpose, an assay based on the activity of the protein of interest is developed to screen for active molecules in a compounds library. The screening hits serve as starting points for the design of molecular probes, which are then used to investigate the behavior of the protein in its environment.

Such analysis involves a three-step process that begins with the development of a suitable phenotypic assay that allows the readout of the desired biological effect. For example, this can be the degree of differentiation of a cell line. Frequently, reporter gene assays are used in this

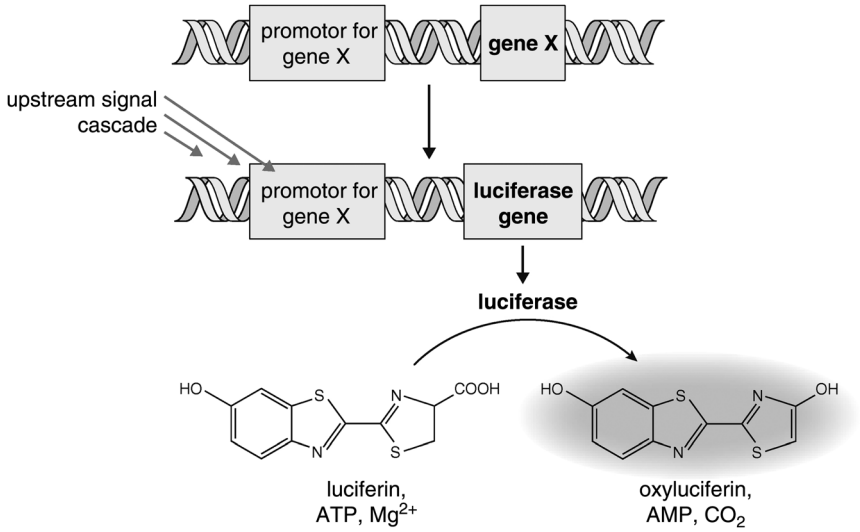


Figure 42.5 Reporter gene assay. The genetic information for the enzyme luciferase is cloned downstream of the promoter of the gene under investigation. The resulting constructs are transfected into suitable cells and, by measuring the luciferase activity, the effect of small organic molecules on the upstream signaling cascade can be read out directly.

context (Figure 42.5). To enable the testing of as many substances as possible for their biological activity in the second step, today assays are miniaturized to microtiter plate format (96, 384, or 1536 wells). After the discovery of compounds, the identification and biological validation of the cellular targets, whose modulation lead to the signal in the assay, is often the greatest challenge. With a new compound, which has been identified as the modulator of the function of a gene product, reverse chemical genetics becomes possible.

42.2.3 The Bump-and-Hole Approach of Chemical Genetics

As we have seen in the previous section, the modulation of protein functions with small organic molecules is a very effective method and the core discipline of chemical biology. The development of molecules, however, that can selectively chemically switch off the function of certain target proteins is very difficult. Inhibitors, for example, often affect other proteins as well as the intended target, particularly those that are structurally or functionally related. Development of a new drug often takes a dozen years and enormous financial resources, which shows that development of a drug is a difficult task. The development of a selective drug without undesirable side effects is impossible (side effects result from interactions with other proteins in many cases). To get around this central problem, at least for the analysis of biological systems and to combine the advantages of small organic molecules (Table 42.1) with the high precision of genetic technology, the bump-and-hole method was developed. This involves extending the binding pocket of the targeted protein by targeted mutagenesis to create a hole that a more sterically demanding ligand, the bump, will bind to selectively, without disturbing the function of the (non-mutated) wild-type or related proteins. The ligand can be an inhibitor, an agonist, or a chemically modified cofactor, in order to, for example, analyze the substrate specificity of the target enzyme in the context of a living cell (Figure 42.6). The bump-and-hole approach was first used to investigate the interaction between the elongation factor EF-TU and the ribosome and has since then been applied to many enzyme classes. Particularly in the laboratory of Kevan Shokat at the University of California, San Francisco, this method has been developed into a powerful chemical genetics approach to determine the function of protein kinases in complex systems. This approach is also called analog sensitive kinase alleles (ASKAs).

ASKA – the Combination of Chemical and Genetic Methods for the Selective Inhibition of Protein Kinases In 1839 scientists had already discovered that the element phosphorus is present in proteins. However, it took another 120 years before the enzymatic transfer of phosphorus to proteins in the form of phosphate groups was discovered and thereby the basis for the understanding of this key post-translational modification.

Today we know that protein kinases transfer the γ-phosphate group of ATP to protein substrates and thereby play key roles in the control of protein functions and the complex

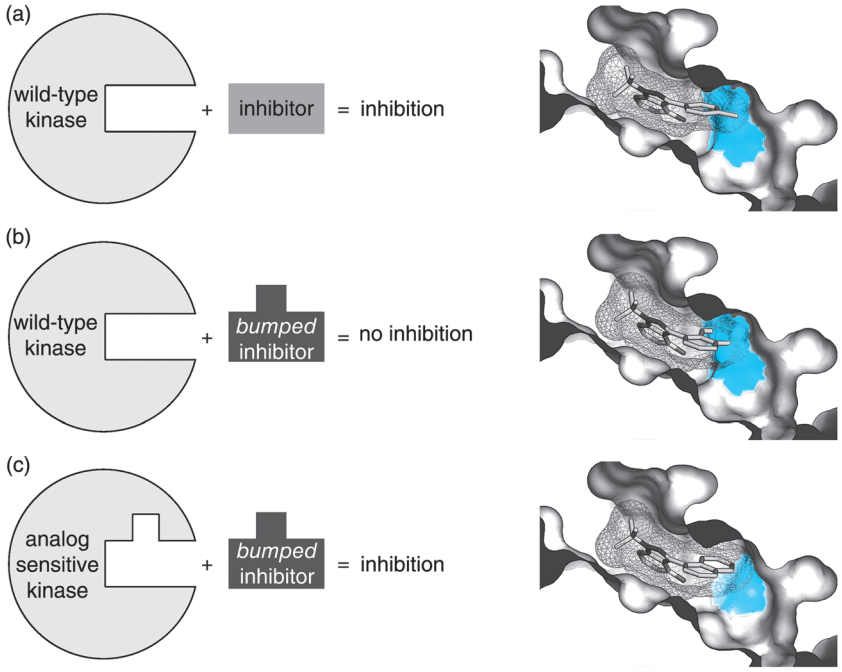


Figure 42.6 The ASKAs (analogue-sensitive kinase alleles) approach creates highly selective kinase/inhibitor pairs (schematic views on left-hand side; structure of the ATP binding pocket in complex with ligand on the right-hand side). (a) Inhibition of a protein kinase by an unspecific, ATP-competitive inhibitor. (b) After modification with a bulky substituent, the inhibitor forms unfavorable interactions with the gatekeeper amino acid (highlighted in blue) in the back of the ATP binding pocket. (c) A point mutation of the gatekeeper residue enlarges the ATP pocket, enabling it to bind the sterically demanding bumped inhibitor.

regulation of signaling pathways. The human genome encodes over 500 kinases and the physiological, and pathophysiological, significance of each member of this enzyme family is of particular interest to current research. However, the complex regulation of these enzymes, as well as their often overlapping substrate specificity, makes the identification of individual kinase activities in biological processes extremely difficult. To answer the question of which kinase is active, at what time point, in which network in the development of an organism and how this activity influences downstream signaling or processes, requires particularly powerful methods. With what we have learned in the previous sections, the answers should be relatively easy to find with the help of perturbation experiments in which kinase inhibitors block the enzymatic transfer of phosphate groups to their substrates and thereby disrupt biological systems. However, this requires extremely selective kinase inhibitors that specifically inhibit only the kinase of interest. The development of potent and selective inhibitors is the central problem of kinase research. Every kinase uses ATP as a cofactor, and most known kinase inhibitors compete directly with ATP for the binding site in the catalytic center of the kinase domain. Since this catalytic domain and, in particular, the ATP binding site are highly conserved, the development of mono-selective inhibitors is an almost impossible task. To study the function of kinases with the help of inhibitors, nevertheless, clever approaches have been developed that combine chemical and genetic methods.

The ATP-binding site of the kinase under study is enlarged by a point mutation such only that a specially developed, sterically demanding inhibitor binds to the kinase and blocks the phospho-transfer. The artificially created complementarity of the ligand and protein structure can reach a surprisingly high level of selectivity, which can be applied to almost any kinase in a number of organisms one to one. This is made possible by the exchange of the gatekeeper, which is a conserved, often large, hydrophobic amino acid located in the ATP binding site. The exchange of this sterically demanding amino acid (phenylalanine, methionine, etc.) for a smaller residue such as alanine or glycine creates additional space in the pocket that is not present in the wild-type kinase.

If one selects a relatively unspecific kinase inhibitor, such as the pyrazolo-pyrimidine PP1 (Figure 42.7) and modifies its structure such that the additional steric needs of the new analogue (NM-PP1) are complementary to the enlarged ATP pocket in the mutated kinase, the mutated (analogue-sensitive) kinase is inhibited, while the wild-type kinase is unaffected (Figure 42.6). By repeated use of the same sterically demanding inhibitor and the mutation of the gatekeeper in the corresponding target kinase, it is relatively simple to generate a number of analogue sensitive kinase alleles and to selectively inhibit them. Remarkably, although the gatekeeper is located in the proximity of the ATP binding site, the exchange usually does not affect the function of kinases. The use of inhibitors that are suitable for ASKAs opens other excellent

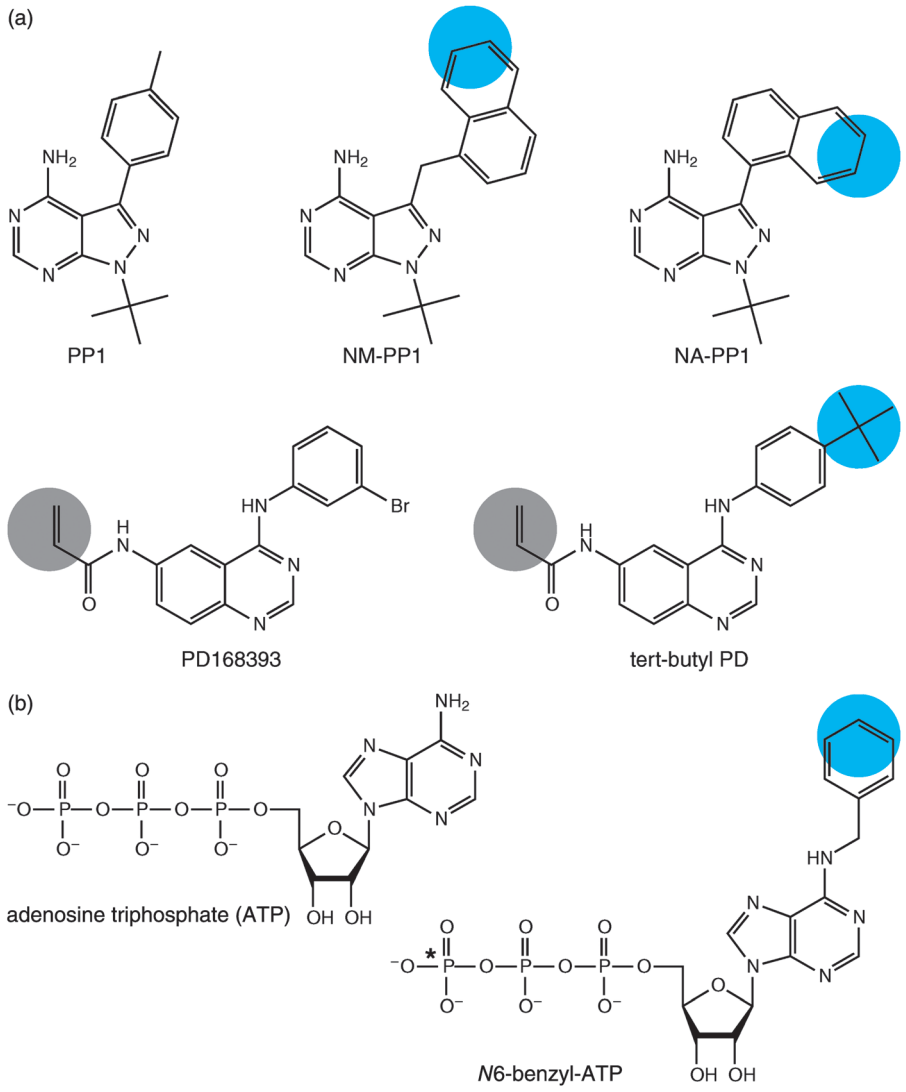


Figure 42.7 Inhibitors and cofactors of analogue-sensitive kinases. (a) Based on classical ATP competitive kinase inhibitors, sterically demanding analogs that are complementary to analogue-sensitive kinases can be generated. The group highlighted in blue prevents binding to the wild-type kinase. A further modification of the inhibitor (e.g., with a Michael acceptor (gray)) allows the covalent modification of cysteines introduced into the target kinase by site-directed mutagenesis. (b) Modified ATP analogues are selectively recognized by analog-sensitive kinases and can be used to transfer radioactively labeled phosphate groups (*, ^{32}P-label).

opportunities for the experimental analysis of the cellular function of kinases. Since the previously mentioned inhibitors NM-PP1 and NA-PP1 are readily taken up from media on cells or can be added to animal feed, they enable a rapid and dose-dependent perturbation of the analogue-sensitive kinases. In addition, the subsequent analysis of the induced phenotype is not limited in any way, since the process is perfomed in an almost native system. Of particular interest is the comparison of changes in gene expression profiles (microarray technologies) that arise from the selective inhibition of analogue-sensitive kinases. The resulting information can contribute to the elucidation of cellular networks, which are regulated by the kinase of interest, or for the identification of new inhibitors.

In the latter case, the expression profile of cells that were treated with the new compound is compared to the profiles of cells treated with the analogue-sensitive kinase's complementary inhibitor. The advantages of ASKA technology relative to purely biological methods can be illustrated in a series of examples.

DNA-Microarray Technology, Chapter 37

Bruton tyrosine kinase (BTK) plays an important role in the immune response. Knock-out experiments that interrupted the function of BTK did not result in a clear phenotype because other kinases in the network associated with BTK compensated for its loss. On the other hand, the speed of the chemical perturbation of the analogue-sensitive variant BTK does not leave the cells sufficient time to compensate the loss of BTK activity. With the use of ASKA-BTK, a much clearer picture of the regulation of cellular processes by BTK was possible. It is often interesting to correlate the strength of the activity of a particular kinase with the observed effect in an animal model. The classical approach to this problem would require the generation of a

large number of mouse lines, each of which expressed a different amount of the kinase of interest. AKSA, on the other hand, only requires a single mouse line and enables the regulation of the kinase activity by the simple adjustment of the dose of the corresponding inhibitor.

The Gatekeeper Residue, an important Amino Acid in the Kinase Domain As mentioned in the previous section, the gatekeeper amino acid is one of the few amino acids near the ATP binding pocket that can be mutated without seriously affecting the activity of the enzyme. In addition, the side chain of the gatekeeper is one of the most important determinants for the development of selective therapeutic kinase inhibitors. While the polarity of the side chain, for example, in the case of threonine, allows the direct interaction with ligands via hydrogen bonds, the size of the residue controls access to the deeper reaches of the ATP binding pocket. It is therefore not surprising that during the treatment of cancer patients with kinase inhibitors clinically relevant resistance mutants emerge at exactly this spot, preventing the binding of the inhibitor. The gatekeeper residue is, therefore, an important amino acid, which, on the one hand, enables innovative approaches to the characterization of biological systems and, on the other hand, presents a great challenge to targeted tumor therapy. The realization that the gatekeeper is a hot spot for the emergence of resistance only came years after the development of the first analogue-sensitive kinase alleles were created in laboratories. However, the principle behind the emergence of resistance by steric blocking can also be used for targeted chemical-biological experiments. To validate a particular kinase as a biologically relevant target of a newly discovered or newly developed inhibitor, the gatekeeper can be used to deliberately introduce larger amino acids at this position and thereby to create an artificial resistance. As a result, the wild type is inhibited but not the mutated kinase.

42.2.4 Identification of Kinase Substrates with ASKA Technology

To understand the biological function of a protein kinase in detail, knowledge about the corresponding substrate proteins is of central importance. Over the course of years methods have been developed that allow the identification of kinase substrate pairs in their biologically relevant context by the use of radioactive ATP (the γ-phosphate group of ATP is marked with the radioactive phosphorus isotope ^{32}P). But the analysis is complicated by the fact that protein kinases often have overlapping substrate specificities and experiments often contain a complex mix of radioactively labeled proteins, which makes the association of a particular kinase with the phosphorylation difficult or even impossible.

The previously mentioned analogue-sensitive methods allow for the facilitation of sterically demanding and radioactively labeled ATP analogues that can be only recognized and enzymatically processed by previously mutated kinases. To adapt the kinase to the changed ATP analogues, the gatekeeper residue in the hinge region of the kinase domain is also mutated to a smaller amino acid (e.g., alanine or glycine). The complementary ATP analogues are chemically modified in the N6 position of the adenine ring system, for example, through the introduction of a benzyl or cyclohexyl group (Figure 42.7b). Correspondingly the γ-phosphate from the modified ATP-analogue is transferred to the appropriate substrates. The use of an inhibitor complementary to the analogue sensitive kinase can be combined with this approach. The changes in the phosphorylation pattern caused by the use of the ASKA kinase and corresponding inhibitor are then analyzed. With the aid of this approach, it was possible to identify several previously unrecognized substrates. Frequently, however, the low copy number of the substrate proteins makes this difficult. To improve the situation, special ATPγS analogues were developed, which are used by analogue-sensitive kinases and transfer a thiophosphate group instead of a phosphate group to the substrate. The special reactivity of the phosphorothioate anion attached to the substrate allows the conversion of *p*-nitrobenzyl mesylate into a phosphothioester. This conjugate is recognized as a hapten by specially created monoclonal antibodies. With this method, not only can the transfer of the phosphate group (thiophosphate group) be followed without the need for radioactivity, it also allows an enrichment of the substrate by immunoprecipitation. With the aid of this approach, it was possible to identify the nucleoporin TRP, which plays an important role in the transport of proteins into the nucleus as a substrate for the MAP kinase Erk2.

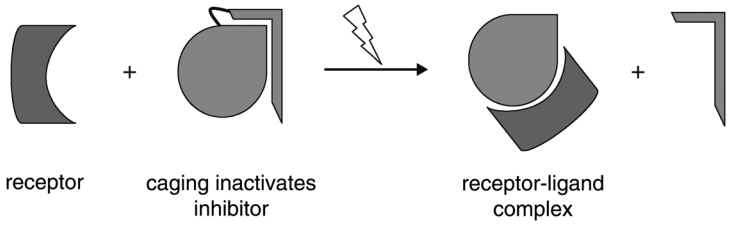
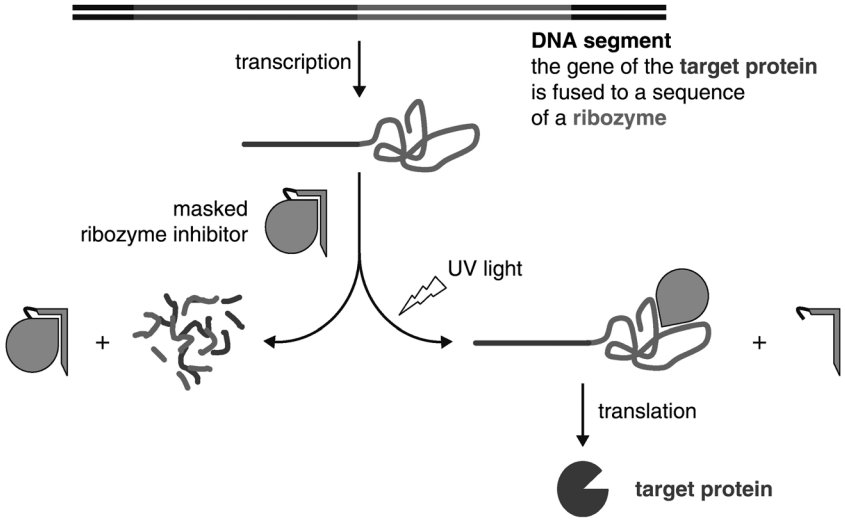


Figure 42.8 Caging gives temporal and spatial control (through external stimuli) over the release of molecular probes. The activity of the probe is initially masked or caged. Exposure to UV light, for example, uncages the probe, which is released to bind to its target structure and modulate its biological function.

42.2.5 Switching Biological Systems on and off with Small Organic Molecules

The presented methods and techniques have shown how protein function can be controlled with the aid of small organic molecules. The expression of target genes can also be modulated in this manner. The best-known example is the Tet system in which expression of the target gene is placed under the control of doxycycline. The tetracycline doxycycline is a broadband antibiotic that binds specifically to the Tet-repressor (TetR), which plays a central role in bacterial resistance to tetracycline. To make the expression of chosen genes controllable by doxycycline, TetR, a doxycycline-inducible DNA-binding unit, can be fused to eukaryotic gene regulation domains. With the aid of this system, selected oncogenes can be brought under the inducible control of doxycycline in transgenic mice, which allows the pathological effects of these genes in the development of cancer to be studied, for example. This elegant method allows the temporal and temporal control of gene expression, which is often interesting for organ-specific expression in animal studies. An alternative is caging, which modifies the chemical probe such that an interaction with the corresponding target structure is impossible and the probe is initially inactive – the biological activity is, in effect, trapped in a cage. Only an external stimulus, like a temperature change, pressure, or irradiation with UV light, breaks open the cage and allows the probe to become active (Figure 42.8).

The activation (uncaging) of a caged probe by photolysis is particularly attractive. In a transparent organism or a cell, even the smallest regions can be addressed under a microscope and the activity of the freed probe can be investigated in an organ, a cell compartment, or a patch of a membrane. For example, a photosensitive doxycycline derivative was used this way for the light-inducible control of expression of target proteins in the brains of mouse embryos. Further approaches to the direct control of gene expression are based on photoactivatable antisense DNA or RNA molecules, which only bind their target structures under the influence of light and thereby cause their degradation in the cells. To make oligonucleotides controllable with light, the photolabile group can be attached either directly to the bases or the phosphate backbone. In principle, caging can also be applied to gain spatial and temporal control of the activity of any molecule – inhibitors, cofactors, or substrates (Figure 42.9). A similar approach involves the



Line of a Ribozyme, Section 38.2

Figure 42.9 Temporal control of the expression of a target protein. The coding sequence of a ribozyme is fused upstream of the genetic information of the target gene. The chosen ribozyme is catalytically active and degrades the coding sequence of the fusion protein after transcription – translation does not take place. Release of a ribozyme inhibitor by light inhibits the catalytic activity of the ribozyme, which enables the transcription and translation of the target protein.

use of molecular probes that can be switched between a biologically active and a biologically inactive form with monochromatic light. Through the reversibility of the isomerization of the probe dependent on the wavelength of the monochromatic light, macromolecules such as ion channels or other receptors become switchable and, therefore, can be investigated conditionally in biological systems.

42.3 Expressed Protein Ligation – Symbiosis of Chemistry and Biology for the Study of Protein Functions

One of the great surprises of early genome analysis was the recognition that the size of a genome is not representative of the complexity of an organism. For example, the genome of the roundworm *Caenorhabditis elegans* with 15 000 genes and that of the fruit fly *Drosophila melanogaster*, with 20 000 genes are only marginally smaller than the genome of modern humans (ca. 25 000 genes). The morphological and functional differences between the organs are fundamentally different from us humans and a fly or worm. If it is not the number of the gene products, what, then, is responsible for the complexity of an organism and life, in general?

Craig Venter, one of the biochemists who was an important participant in decoding the human genome, stated the following in one of his highly regarded articles in the journal *Science*: "The finding that the human genome contains fewer genes than previously predicted might be compensated for by combinatorial diversity generated at the level of post-translational modification of proteins." Post-translational modifications of proteins are specific, enzyme-catalyzed covalent modifications that change the information content of proteins. Indeed, post-translational modifications play a central role in the regulation of all metabolic processes. Nature can thereby control the function of proteins by, for example, a change in stability, charge, cellular localization, three-dimensional structure, or interaction with other molecules. In cells post-translational modifications, such as phosphorylation, conjugation with lipids, or glycosylation, are usually subject to a finely tuned, reversible exchange that allows a previously introduced modification to be removed again, returning the system to its beginning state. Regulatory mechanisms based on post-translational modifications are highly dynamic, which makes them interesting for the elucidation of biological questions, but also difficult to handle at the same time. Often, the modified protein preparations necessary for detailed investigation of the function cannot be obtained by purely biological methods, such as targeted mutagenesis or recombinant protein expression. The development and application of combined chemo-biological techniques for the chemoselective modification of proteins have proven to be a superb tool for the study of protein function on the molecular level and represents another important focus of chemical biology.

42.3.1 Analysis of Lipid-Modified Proteins

The originally purely synthetic method of native chemical ligation (NCL) has proven to be extremely useful for protein chemistry. This method allows the synthesis of large peptides by the condensation of peptide fragments, consisting of a C-terminal thioester and an N-terminal cysteine. The semi-synthetic version of NCL, known as expressed protein ligation (EPL), combines chemical synthesis with biological techniques. This method allows the fusion of synthetically manufactured peptides with recombinantly synthesized proteins. EPL allows the site-specific modification of proteins with a large number of probes like fluorophores, spin labels, stable isotopes, unnatural amino acids, or post-translational modifications and has already been applied successfully to a large number of questions of protein design. A previously selected reporter group (e.g., fluorophore or unnatural amino acid) is chemically built into the peptide. The ligation of the peptide with the recombinant protein fragment allows, for example, structural biological investigations of the previously mentioned post-translationally modified proteins (Figure 42.10).

The example shown in Figure 42.10 demonstrates that EPL could be used to obtain preparative amounts of mono- and diprenylated variants of the Rab-GTPase Ypt1, which

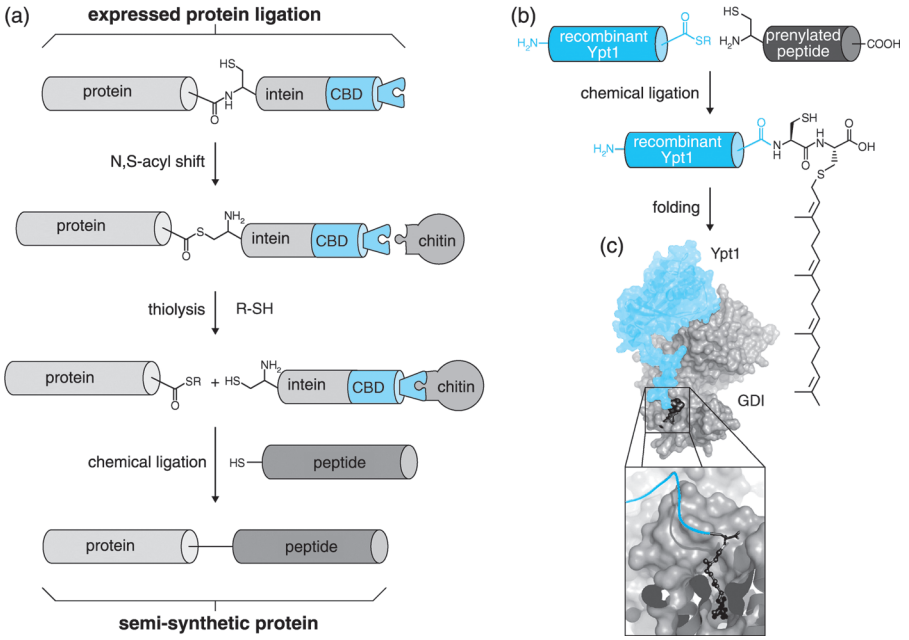


Figure 42.10 *Expressed protein ligation* (EPL). (a) A protein splicing element (intein) is recombinantly overexpressed along with an affinity tag such as the chitin-binding domain (CBD) fused with the C-terminus of the target protein. The N-terminal cysteine of the intein domain initiates an N,S-acyl shift, after which exogenous semi-synthetic thiols cleave the intein by thiolysis of the recombinant thioester of the target protein. The final, semi-synthetic target protein is formed by the reaction of the C-terminal thioester with the N-terminal cysteine of the synthetic peptide through a ligation reaction. Chitin beads allow the effective separation and purification of the formed construct. Commercially available systems have been developed for the intein-mediated purification of immobilized chitin. There are numerous examples of biological questions that could be addressed by use of EPL. One example is the X-ray structure of the monoprenylated Ypt:RabGDI complex (PDB-code 1ukv). (b) EPL made the isolation of preparative amounts of mono- and diprenylated variants of the Rab-GTPase Ypt1 possible. The prenyl-modified and synthetically produced dipeptide (gray) is ligated with the C-terminus of the recombinantly obtained Ypt1 (blue). (c) After successful folding, the crystallization and structure determination of the semi-synthetically modified Ypt1 in complex with its physiological modulator RabGDP-dissociation inhibitor (RabGDI) was possible for the first time. Rab proteins are important regulators of vesicular membrane transport and mediate numerous events, such as the docking and fusion of membranes as well as their intracellular mobility. Post-translational modifications like prenylation are therefore essential for protein function. The X-ray structure of the Ypt1:RabGDI complex showed a conformational change of RabGDI induced by binding of the prenylated Ypt1 and an associated formation of a hydrophobic binding pocket that accepts the prenyl side chain of Ypt1. In the uncomplexed state the prenyl moiety is anchored in the plasma membrane. The illustration shows the surfaces and secondary structure of RabGDI (light gray) bound to Ypt (blue) with the prenyl moiety (dark gray). The close up shows the binding of RabGDi to the hydrophobic pocket formed by the prenylated Ypt1. Source: from Rauh, D., and Waldmann, H. (2007) *Angew. Chem., Int. Ed. Engl.*, **46**, 826–829. With permission, Copyright © 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

enabled their crystallization and structure determination in a complex with their physiological modulator RabGDP-dissociation inhibitor (RabGDI). Rab-proteins are GTPases of the Ras superfamily and have a central role as regulators of vesicular membrane transport. They modulate numerous events such as the docking and fusion of membranes and their intracellular mobility. Post-translational modifications are essential for protein function and the course of physiological processes. The X-ray structure of the Ypt1:RabGDI complex showed a conformation change in RabGDI induced by the binding of the prenylated Ypt1 and an associated formation of a hydrophobic binding pocket, which accepts the prenyl side chain of Ypt1. In the uncomplexed state, the prenyl moiety is anchored in the plasma membrane of the cell and therefore sequestered. The group of Roger Goody at the Max-Planck-Institute in Dortmund was the first to determine the position of the prenyl binding site and led to the elucidation of the molecular mechanism of the insertion and removal of Rab proteins through effectors like Rab or GDI. Interestingly, mutations in RabGDI can cause mental retardation in humans. By the use of prenylated proteins obtained with EPL, many comprehensive biophysical and structure biological studies could be carried out which showed that mutated Rab proteins like prenylated Ypt are harder to remove from the membrane. Thus, a disruption of membrane transport is the ultimate molecular cause of this genetic disease.

42.3.2 Analysis of Phosphorylated Proteins

As discussed in the previous sections, the enzymatic transfer of phosphate groups to the side chains of serine, threonine, and tyrosine is a central modification of proteins and is essential for the maintenance of almost all biological systems. On the cellular level, kinases oppose the action of phosphatases, which hydrolytically remove the phosphate groups. The orchestrated balance between phosphorylation and dephosphorylation forms the basis of signal transduction that allows exchange of information between different compartments of the cell, for example. A detailed biological understanding of these complex processes as well as the recognition that the dysregulation of the cooperation between the two enzyme classes can be causal to the genesis and progression of diseases like cancer, auto-immune diseases, diabetes, or neurological defects has led these signal transducing proteins to be considered promising target proteins in modern drug research. However, the detailed biochemical and biological characterization of the influence of phosphorylation patterns on the function of proteins has proven to be very difficult due to the tendency towards dephosphorylation, particularly under physiological conditions. These difficulties can be elegantly avoided by the semi-synthetic incorporation of, for example, phosphomethylene-L-phenylalanine (Pmp) as a well-characterized, non-hydrolysable phosphotyrosine mimic for the production of homogenous protein preparations. This phosphonate imitates the functional phosphorylation on the essential positions in the protein under investigation. The use of EPL techniques can be applied to the phosphorylation of proteins *in vivo* by the microinjection of semi-synthetically produced proteins equipped with non-hydrolysable phosphotyrosines into living cells. Besides the examples of semi-synthetic methods for the study of the modification of lipids and phosphorylation, a series of other ligation techniques have been developed, which, among others things, allow the transfer of complicated glycosylation patterns to proteins.

Posttranslational Modifications, Chapter 25

42.3.3 Conditional Protein Splicing

The utility of protein ligation for the study of protein functions is apparent and therefore there is a great deal of interest in the development of minimally invasive techniques that allow an analysis of the target proteins in the complex environment of all proteins *in vivo*. One of these methods is conditional protein splicing. This is an *in vivo* technique that makes possible the reconstitution of two inactive protein fragments into the functionally active target protein under the control of a cell-permeable, small molecule ligand (Figure 42.11). The methods of conditional protein splicing are based on the fusion of inactive domains of the target protein with the two protein fragments of an excised inteins. The two intein halves are fused, in turn, either to the rapamycin receptor or with the rapamycin binding domain. Under the influence of rapamycin, the two rapamycin-binding proteins dimerize. The resulting complex formation of the intein fragments initiates the autocatalytic splice reaction to reconstitute the functional and active target protein. This elegant method enables the activation of protein functions under the

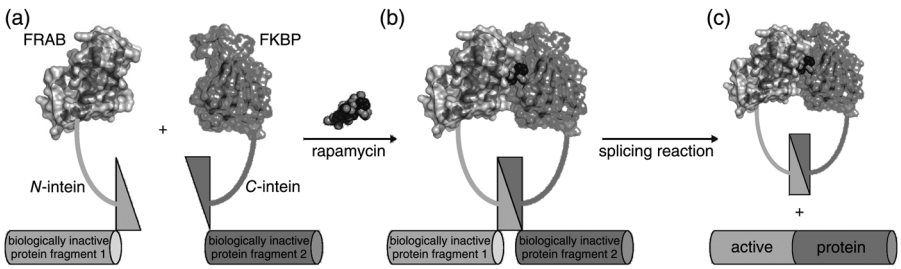


Figure 42.11 Conditional protein splicing. As an important development of chemical-biologic ligation techniques, this method allows protein functions to be turned on by small organic molecules in cellular systems. (a) The biologically inactive protein fragments (1 and 2) of the target protein to be reconstituted are cloned as fusions of cleaved inteins (*N*-intein and *C*-intein) and rapamycin-binding proteins (FRAB und FKBP). The fused constructs are expressed in cells. (b) Under the control of the small molecule, cell-permeable, natural substance rapamycin, the rapamycin-binding domains of the FRAB and FKBP interact, which leads to dimerization of the two constructs and results in the reconstitution of the intein. (c) The functional intein initiates the splice reaction of both protein fragments 1 and 2 to the biologically active target protein RabGDI. Source: from Rauh, D., and Waldmann, H. (2007) *Angew. Chem., Int. Ed. Engl.*, **46**, 826–829. With permission, Copyright © 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

control of small organic molecules in *in vivo* systems and represents an example of the development of ligation techniques in the field of chemical genetics.

The ligation methods briefly described here are innovative tools for the study of protein functions in living cells and make it possible to better understand and follow the function of proteins in the dynamic and variable environment of biological processes.

Further Reading

Baker, A.S., and Deiters, A. (2014) Optical control of protein function through unnatural amino acid mutagenesis and other optogenetic approaches. *ACS Chem. Biol.*, **9**, 1398–1407.

Bishop, A.C., Ubersax, J.A., Petsch, D.T., Matheos, D.P., Gray, N.S., Blethrow, J., Shimizu, E., Tsien, J.Z., Schultz, P.G., Rose, M.D., Wood, J.L., Morgan, D.O., and Shokat, K.M. (2000) A chemical switch for inhibitor-sensitive alleles of any protein kinase. *Nature*, **407**, 395–401.

Dar, A.C. and Shokat, K.M. (2011) The evolution of protein kinase inhibitors from antagonists to agonists of cellular signaling. *Annu. Rev. Biochem.*, **80**, 769–795.

Hasson, S.A. and Inglese, J. (2013) Innovation in academic chemical screening: filling the gaps in chemical biology. *Curr. Opin. Chem. Biol.*, **17**, 329–338.

O'Connor, C.J., Laraia, L., and Spring, D.R. (2011) Chemical genetics. *Chem. Soc. Rev.*, **40**, 4332–4345.

Rak, A., Pylypenko, O., Durek, T., Watzke, A., Kushnir, S., Brunsveld, L., Waldmann, H., Goody, R.S., and Alexandrov, K. (2003) Structure of Rab GDP-dissociation inhibitor in complex with prenylated YPT1 GTPase. *Science*, **302**, 646–650.

Waldmann, H. and Janning, P. (2014) *Concepts and Case Studies in Chemical Biology*, Wiley-VCH Verlag GmbH, Weinheim.

Toponome Analysis

43

"Life is Spatial"

The hierarchy of cellular functionalities consists of at least four functional levels: genome, transcriptome, proteome, and toponome. They exist as interacting spatial systems and subsystems, whose proper functions require correct topologically determined compositions of their molecular components. Moreover there is no doubt that a biological function not only depends on the amount of participating molecules, but in particular the local context – the molecular neighborhood inside the cellular or tissue structures plays an important role. For example, it has been shown that the relative concentration and differential relative spatial order of more than 20 proteins at the cell surface membrane enable a tumor cell to enter an explore state preceding and inducing the migratory state. Furthermore, similar investigations in clinical samples have shown that co-mapping of a large number of molecular components was key for the investigator to find a hierarchy of molecular networks and predict a disease specific target molecule whose downregulation leads to clinical effects, as shown for amyotrophic lateral sclerosis. These examples have shown that the *in vivo/in situ* topology of large molecular systems with their inherent topological hierarchies, controlled by lead proteins, is important for the finding of clinically relevant target molecules. If this topological context is destroyed by tissue homogenization or isolation of components the spatial context of biomolecules, the toponome, is lost. Hence toponome *in situ* technologies are needed when it comes to the spatial analysis of large molecular systems (toponome) driving cellular function or dysfunction in disease – the so-called disease robustness networks (Schubert 1995; 2015).

Toponome By the term toponome we understand the spatial arrangement of the molecular networks in a cell or a tissue, for.mally described by a colocalisation and anti-colocalisation code.

Toponomics addresses the structure and (dys)function of molecular systems *in situ.*

Toponomics Analysis of the modes and rules of the quantitative combinatorial arrangements of the molecular components of a biological structure or system under defined conditions.

In this chapter the basic methods of toponome analysis are described, which have already been shown to provide new insight into the operations and functions of large molecular systems inside cells and tissues with their predictive power in the field of diagnostics and clinical medicine.

43.1 Antibody Based Toponome Analysis using Imaging Cycler Microscopy (ICM)

Walter Schubert[1,2]

[1]*Molecular Pattern Recognition Research Group, University of Magdeburg, Leipziger Straße 44, 39120 Magdeburg, Germany*
[2]*Human Toponome Project, TopoNomos Ltd., Margaretenstraße 20, 81373 Munich, Germany*

Biological systems Biological systems imaged in toponomics possess emergent (also called relational or systemic) properties that are possessed only by the system as a whole and not by any isolated part of the system.

Considering the definition of the term toponome, we anticipate that most cellular functionalities are based on the interaction of proteins and other molecular components of the cell. These

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

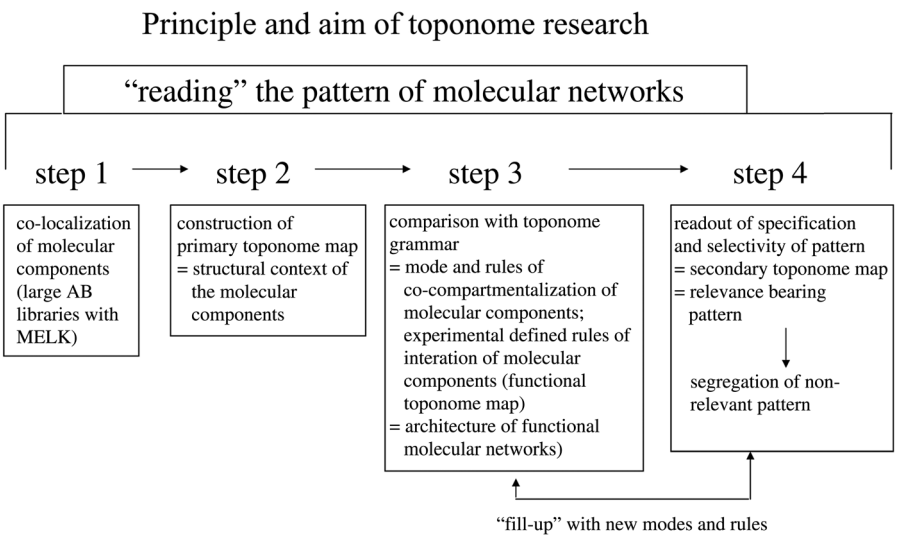


Figure 43.1 Scheme: toponome reading process with ICM leading to the spatial decoding of molecular networks of a cell or a tissue. Using pattern recognition, topological *in situ* mapping, or experimental and/or clinical analysis (clinical trial) the rules of the toponome "grammar" are found, and then integrated as functional knowledge into next step new reading processes (progressive decoding of the spatial molecular networks in health and disease – a pattern cognition like procedure).

Imaging cycler microscope (ICM) The fundamental automated labeling principle overcoming the spectral fluorescence resolution limit to detect a random number of molecular components in the identical morphologically intact biological structure (subcellular, cell or tissue) in a single experiment. Stages of development included the multi-epitope-ligand cartography (MELK) and toponome imaging system (TIS).

interactions can involve more or less strong direct physical crosstalk of proteins or can be based on indirect crosstalk (e.g., by means of diffusible molecules). In any case, every element of such interactions (protein or other molecular component) must be at the right time point at the right concentration at the right location inside a cell or on an extracellular matrix so that a concrete molecular network can be formed and operationally active. Hence, molecular networks are characterized by a specific spatial context of their single elements: every molecular network is based on a highly non-random topology of its molecular elements. As with every system, large molecular systems *in vivo* have intrinsic relational (emergent) properties, which are only possessed by the systems a whole and not by any of its isolated parts.

In other words, the topology of these systems is the direct expression of the molecular compartmentalization rules of the cell or the tissue. Toponomics aims to analyze and map these co-compartmentalization and topological association rules of molecular components as well as their functional network architecture in morphologically intact fixed cells or tissue sections. For this purpose the fully automated imaging cycler technology is applied: with the aid of large tag libraries (mostly antibody libraries) a large number of molecular components is co-localized and visualized in the direct context of cell functions. These data serve to construct functional toponome maps. Progressively, such toponome maps reveal the modes and rules of functional constellations of the molecular components of cells and tissues that are assembled within a so-called toponome grammar (Figure 43.1). While the basic principles of this technology were developed in the late 1980s, only the development of highly automated imaging cycler microscopes and powerful software in the last 20 years has opened up the possibility of systematically establishing the research field of toponomics.

43.1.1 Concept of the Protein Toponome

Toponome map Structural representation of toponome modes (CMPs, CMP groups, CMP motifs) of biological structures (cell or tissue), which are specifically expressed in certain biological regions (e.g., single data point – pixel or voxel – a subcellular compartment or partial compartment, in a whole cell or whole tissue compartment).

Given that most proteins inside and on the surface of cells are organized as networks exerting the different cellular functionalities we also can assume that these proteins are not stochastically distributed but in terms of both time and space are highly organized. This is analogous to written language, in which letters are assembled as words and sentences; following syntactic and semantic rules, the proteins, which are organized as networks, are also topologically precisely determined in every cell. The cell itself thereby functions as a protein pattern formation apparatus, or can be seen as a protein co-compartmentalization machine. As described in the introduction most of the molecular components of the cell or a tissue, including proteins, must be at the right place at the right concentration in the cell to interact with other proteins following similar rules. It follows that every given cellular functionality can be detected as a specific contextual protein pattern directly in the corresponding biological structure and, further, that blocking of the molecular component(s) hierarchically and topologically controlling this entire

structure will result in a disassembly of that structure and loss of the corresponding (dys) functionality. One important task is to find by toponomic analyses the most relevant biomolecule controlling the overall structure of the corresponding molecular systems in a disease, block it therapeutically, and thereby successfully interfere with the disease. The entirety, or fractions, of all protein networks are termed the toponome – a purely descriptive term derived from the ancient Greek nouns *topos* (τοποζ) (place) and *nomos* (νομοζ) (law). This term simply says that the natural *in situ* protein network code is topologically determined, and as such represents a spatially determined functional code exerting the cellular functionalities. The corresponding spatial data sets are amenable to a quantitative, mathematical analyses and combinatorial geometric measures (e.g., for functional genomics), with the interesting perspective that, technically, a direct alignment of next-generation genome sequencing and toponome mapping is in the pipeline. To identify protein networks directly in a given biological structure (e.g., inside a single cell) a microscopic "reading technology" is needed, which must fulfil the following conditions (as provided by imaging cycler technology, below):

- A large, quasi-random number of distinct molecular components (e.g., proteins) must be comappable independently from each other inside the same structure in a single experiment.
- The resulting combinatorial molecular patterns that specifically characterize the corresponding cell type or corresponding cell function (or dysfunction) or a tissue must be visualizable as a whole, and also in selected fractions of the whole, by applying an appropriate interrogation algorithm.
- The biological meaning of these patterns must not necessarily be understood immediately, but can serve as an important specific toponome feature (e.g., a disease process), and should be interrogatable functionally, by aligning an actually found pattern with a memory store that contains the modes and rules of topological protein associations of the cell (toponome grammar) found in preceding toponome mappings (Figure 43.1).

The driving force is a vision of the complete grammar of the protein networks of the cells in human, animal, or plant tissues. While the imaging cycler technology fulfils the criteria and principles of co-localization and anti-colocalisation of an extremely large number of different molecular components (Figures 43.2–43.4) and is capable of visualizing directly these comapped patterns inside or on the surface of a given cell (Figures 43.4e and 43.2) or in a tissue section (Figures 43.3 and 43.4a), the future challenge is to gain insight into the precise spatial coding function of protein networks on a proteome-wide scale to discover all essential topological protein associations. For this purpose toponome mapping and biological experiments must be linked. This has been shown for amyotrophic lateral sclerosis (ALS) clinically, and for tumor cells experimentally. In both cases it was shown that blocking the hierarchical lead proteins of large protein networks leads to a halt in disease progression and loss of the ability to migrate, respectively. Hence in both cases disease robustness networks were found that could be broken by inhibiting the lead proteins. Basic cancer specific lead protein-connected toponomes were found in prostate cancer and in colon cancer.

43.1.2 Imaging Cycler Robots: Fundament of a Toponome Reading Technology

Chemical Modifications of Proteins, FITC, Section 6.2.1

Toponome reading requires highly automated procedures. It was established as a system of so-called image cycler workstations (for its set up and operations see www.huto.toposnomos.com). These robots consist of a multi-pipetting unit, a fluorescence microscope, and a CCD camera. The robots can be fully programmed to incubate large tag libraries (e.g., antibodies) on fixed cells and tissue sections to co-localize their molecular components in one single experiment. A large series of experiments over 20 years established that the most stable procedure is as follows: Each tag is conjugated to the same fluorochrome (e.g., FITC, fluorescein isothiocyanate) and is applied and incubated sequentially on the biological sample, which is placed on the object table of the robot microscope. Every labeling step is followed by an imaging step and then by a bleaching step, so that a large number of iterative labeling–imaging–bleaching rounds can be performed to overcome the spectral resolution limit in fluorescence microscopy. The result of this cyclical process is a two- or three-dimensional combinatorial protein fingerprint at

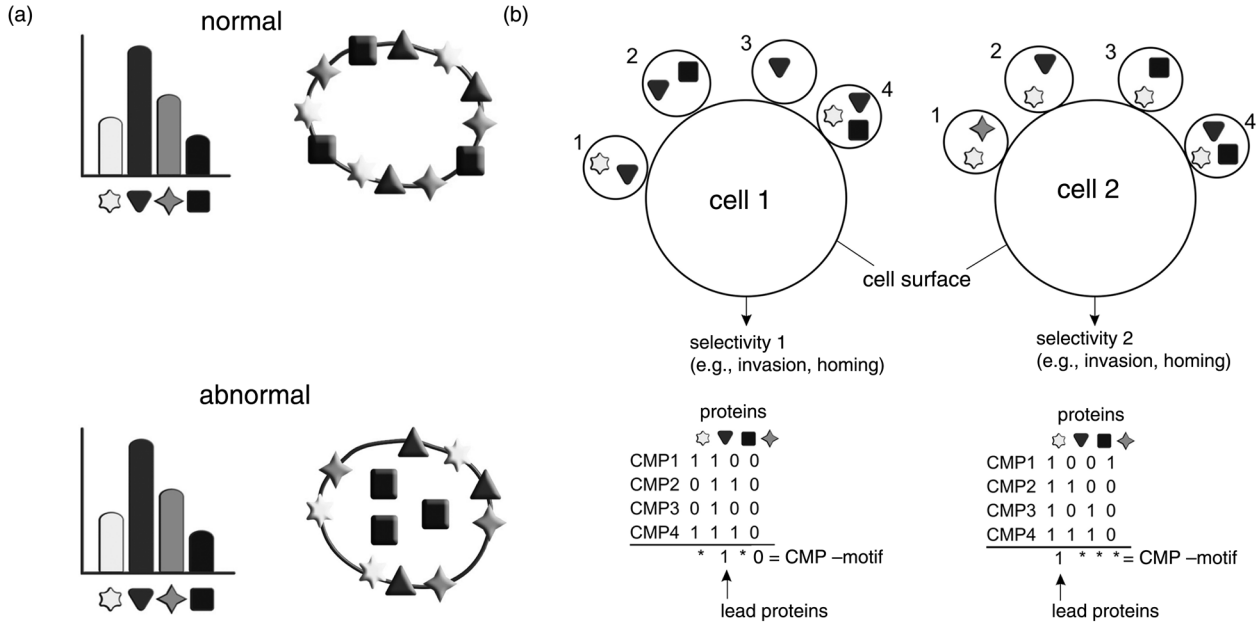


Figure 43.2 Schematic illustration of the toponome theory confirmed experimentally and clinically. Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nat Biotechnol.*, **24** (10), 1270–1278. Copyright © 2006, Rights Managed by Nature Publishing Group; Schubert, W. Advances in toponomics drug discovery: Imaging cycler microscopy (ICM) correctly predicts a therapy method of amyotrophic lateral sclerosis (ALS) Cytometry Part A; 2015, 87A: 696-703. doi:10.1002/cyto.a.22671. (a) Distinct spatial protein toponome patterns *in situ* specifically designate different cells (on the right-hand side). By extraction of the corresponding proteins and their annotation as *ex vivo* protein profiles (left-hand side) the natural *in situ/in vivo* differences between the cellular states "normal" and "abnormal" can no longer be detected because the concentration of these proteins is identical in these two cells. This can be the source of fundamental errors in interpretation and downstream conclusions (e.g., in therapeutic management). (b) Two cells express distinct selectivities on their cell surfaces by differentially combining the identical proteins as higher order units. These different local protein constellations (1–4) can be: (i) a simple co-existence of these proteins without any direct cross interaction, (ii) a non-direct interaction via other not co-localized biomolecules, or (iii) a direct physical interaction with weak or strong binding forces. Together, and independently of these alternative functional constellations, all four distinct local constellations (1–4) are clear cut multimolecular domains (MMDs), which – in the present case – encipher high selectivity for cell–cell or cell–matrix interaction and – most importantly – can couple these differential MMD information units with corresponding MMD specific activation of intracellular mechanisms. Hence, these four domains "encipher" specific supramolecular functionalities (exclusion principle), which together form the ground structure of a functional molecular network. By using imaging cycler microscopes (ICMs), such networks are detectable as specific toponome patterns. ICM generates multi-dimensional vectors of these different protein ensembles that can be collected as toponome lists of different CMPs (combinatorial molecular phenotypes) (b, bottom). Every CMP is a geometric object (one or several data points in x/y or x/y/z). All CMPs together reveal a toponome map of a cell or a tissue. The different CMPs (b, bottom) express both distinct and common features in the two cells, which together reveal distinct CMP motifs. These motifs represent the higher order feature of distinct functional codings of the cell surface.

each subcellular data point in a corresponding cell or tissue sample (scheme in Figure 43.5) (biological/clinical examples in Figures 43.3 and 43.4e, respectively). The resulting high-dimensional context of combinatorial protein information, which would be lost in cell homogenates or lists of proteins, supports the conclusion that protein networks are spatially determined functional units in every cell (different cell functions = different protein networks = unique combinatorial cellular protein pattern). The scheme in Figure 43.2 shows that single normal and abnormal cells are characterized by specific relative locations of the single proteins as a cellular fingerprint, which specifically characterizes these cells. Sometimes these features have a periodical order. If, however, these cells are homogenized the resulting quantitative profiles of the corresponding proteins no longer show any difference (Figure 43.2a, left-hand side). These latter *ex vivo* profiles (without their corresponding spatial combinatorial topologies) might prompt us to conclude that these identified proteins have no relevant function in the corresponding disease, because their abundance is not changed. However, the co-mapping of the same proteins leads to the contrary conclusion (Figure 43.2a, right-hand side). Such constellations have been shown and published for tumor biology and singled out in several comprehensive reviews. The current observations made with ICM (image cycler microscopy) support the working hypothesis that every cell contains a quasi-infinite "toponome space" to spatially encipher the myriad of different cellular functionalities. The combinatorial patterns on a large scale, which are generated by cells *in vivo,* are, as expected, highly restrictive and non-

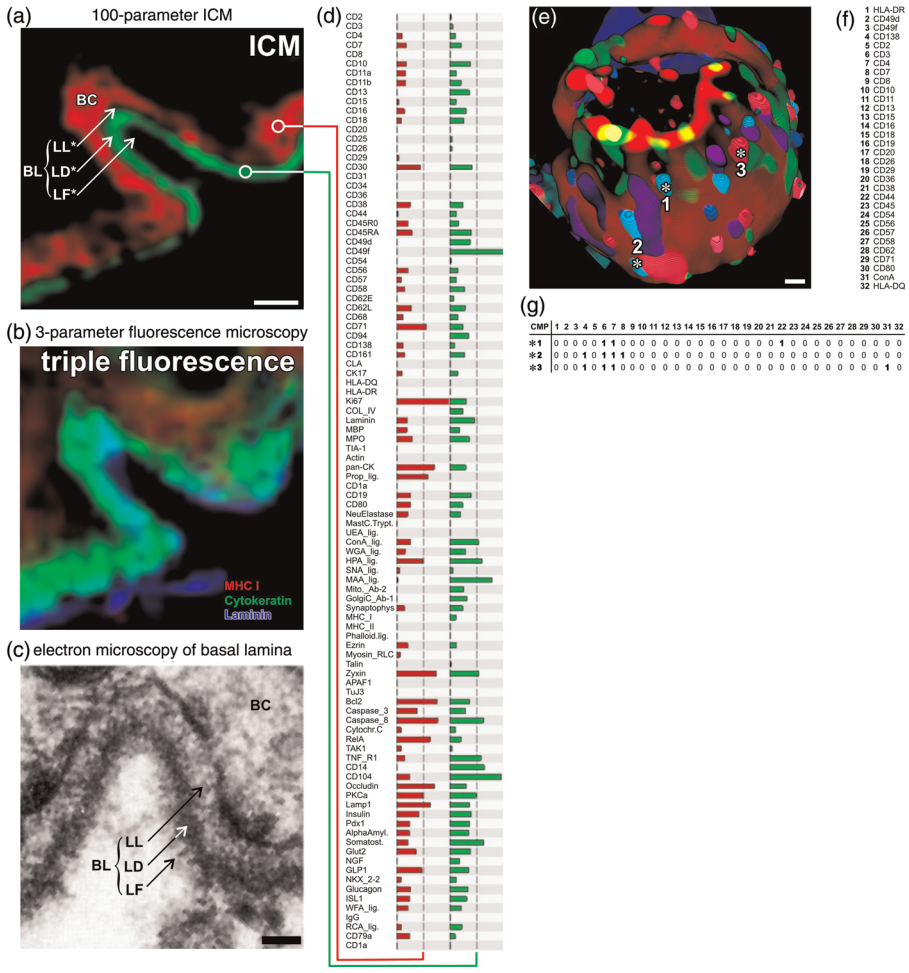


Figure 43.3 Functional super resolution of large molecular networks. (a)–(d) Dermo-epithelial junction in human tissue. Imaging cycler microscopy based discovery of molecular networks *in situ*. (a) Direct real-time protein profiling in 100-dimensional ICM data set using an algorithm based on the similarity mapping approach (Dress, A.W.M., Lokot, T., Pustyl'nikov, L.D., and Schubert, W. (2005) Poisson numbers and Poisson distributions in subset surprisology. *Ann Comb.*, **8** (4), 473–485 (Copyright © 2004, Birkhäuser Verlag, Basel); Dress, A.W.M., Lokot, T., Schubert, W., and Serocka, P. (2008) Two theorems about similarity maps. *Ann. Comb.*, **12** (3), 279–290 (Copyright © 2008, Birkhäuser); Schubert, W., Gieseler, A., Krusche, A., Serocka, P., and Hillert, R. (2012) Next-generation biomarkers based on 100-parameter functional super-resolution microscopy TIS. *Nat. Biotechnol.*, **29** (5), 599–610 (Copyright © 2011 Elsevier B.V. All rights reserved)). Each data point has a power of combinatorial molecular discrimination (PCMD) of 256^{100}. Note: sharp images at the junctional area discriminating between lamina fibroreticularis (LF), lamina densa (LD, green profile in (d)), lamina lucida (LL), and the basal keratinocyte layer (BC, red profile in (d)), as known from transmission electron microscopy (c). (b) Same area as in (a), displaying traditional triple fluorescence imaging. (e) Three-dimensional ICM imaging of distinct 32-component multiprotein complexes on the cell surface of a blood T-lymphocyte (Friedenberger, M., Bode, M., Krusche, A., and Schubert, W. (2007) Fluorescence detection of protein clusters in individual cells and tissue sections by using toponome imaging system: sample preparation and measuring procedures. *Nat. Protocol.*, **2** (9), 2285–2294 (Copyright © 2007, Rights Managed by Nature Publishing Group)). Multiprotein complexes are composed of differential combination of 32 proteins/glycotopes listed in (f). (g) Examples are marked with asterisks (number 1–3) and detailed as combinatorial molecular phenotypes (CMPs) with proteins present (1) or absent (0) together characterized as individual CMPs. Bars: 10 μm (a, b), 50 nm (c), and 1 μm (e).

random. From a pure mathematical point of view the gain of information using the ICM technology, as compared to other methods, is significant. For example, given the toponome acquisition as shown in Figure 43.3 the co-mapping of 100 biomolecules in 1 million data points allows for the simultaneous discrimination of any biologically expressed combination out of 100 biomolecules per data point, when a non-threshold based approach termed similarity mapping is applied on the 100-cycles. Hence the power of combinatorial molecular discrimination (PCMD) per data point is 256^{100} (using an 8-bit CCD camera) or $65\,536^{100}$ (using a 16 bit CCD camera). Compared to *ex vivo* protein profiling the gain of information in this example (using the 16 bit CCD) is $65\,536^{100} \times n$ million data points per image. The information content hence increases exponentially using ICM technology. If the number of co-mapped biomolecules exceeds 100, the information content increases and determines corresponding "reading frames." A specific example is shown in Figure 43.3a–d illustrating the dermo-epithelial junction in a histological skin tissue section analyzed by 100-component molecular profiling: The three layers of the basal lamina, hitherto only seen in electron microscopic images (Figure 43.3c), can be clearly distinguished by ICM (Figure 43.3a), but not by traditional three-parameter fluorescence (Figure 43.3b). By using deconvolution algorithms it is possible to reveal many distinct optical levels with a high voxel resolution so that all three-dimensional data points together reveal the three-dimensional reality of protein network assemblies of many multiprotein complexes at a time (Figure 43.3e, f, g). These measurements can be extended extremely by parallel measuring of cell or tissue arrays. Multiple ICM robots can cooperate strategically to measure very large numbers of probes as a high-throughput/high-content screening procedure. By using optical tools (e.g., navigation algorithms) a complete protein network in every cell can be visualized and explored (walking through large toponome fractions with subcellular resolution). As substantiated by power law, the supramolecular order of

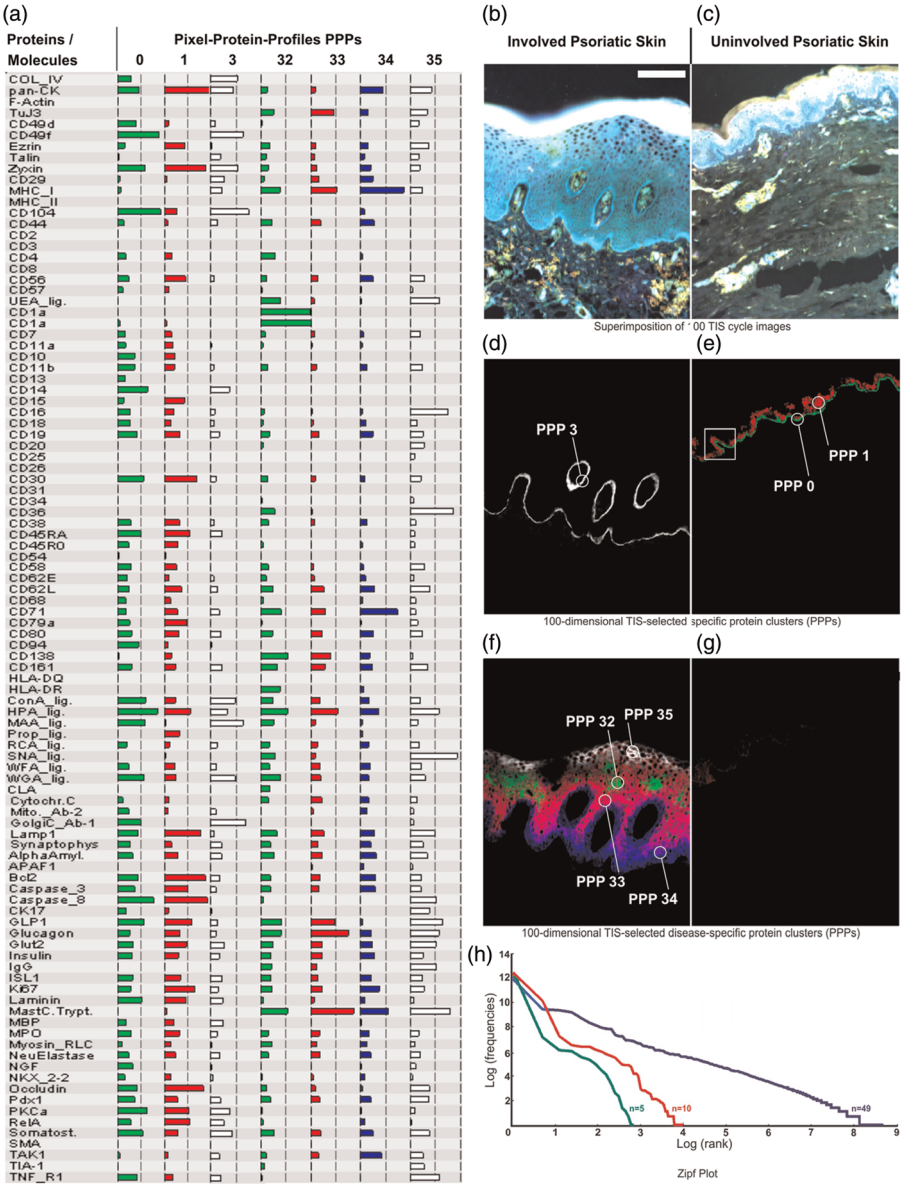


Figure 43.4 Discovery of disease specific 100-dimensional protein profiles simultaneously and in real-time in morphology intact tissue sections at a PCMD of 256^{100} per pixel, exemplified in human skin. (a) List of 100 co-mapped biomolecules and selected protein profiles (PPPs: 0, 1, 3, 32–35) specific for diseased (d, f) and normal skin (e, g). (b, c) Diseased (b) and normal skin (c) are highlighted by pseudo-coloring as histological stain for morphological orientation. Note that, by moving the cursor over the pixels, the software directly recognizes in real-time which protein profiles are specific for the diseased (d, f) or normal skin (e, g). For example, pixel protein profiles (PPP) with numbers 0 and 1 are specific for the normal skin (e), and PPP 3, as well as PPPs 32–35 (d, f, respectively) are specific for the diseased skin. Note: For many similar applications at real-time see webpage of the human toponome (HUTO) project (www.huto.toposnomos.com). (h) Power law (Zipf's law) substantiates highly organized protein systems, as seen in (d)–(g). If 49 molecules are co-mapped Zipf's law applies (blue line), but does not apply if less than 15 molecules are co-mapped (red and green lines). This is revealed by plotting the log–log-relationship of thousands of distinct protein assemblies in toponome data sets (Schubert, W. *et al.* (2006) Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nat. Biotechnol.*, **24** (10), 1270–1278 (Copyright © 2006, Rights Managed by Nature Publishing Group). Bar: 100 μm (b)–(g).

molecular systems is revealed by co-mapping more than 20 molecules, but not by co-mapping lower numbers (Figure 43.4h). Since finding critical lead proteins that control molecular networks depends on the discovery of supramolecular order (see below) this fact is essential for any drug target discovery study.

A topologically determined ensemble of proteins that exerts a given cell function is defined as a protein network motif (combinatorial molecular phenotype motif; CMP-motif). In addition, primary ICM datasets always contain the precise fluorescence intensity at every given data point for every single co-mapped molecular component. In routine work, an effective procedure has been to map every single molecular component relative to a threshold intensity, as present or absent (present = 1; absent = 0; 1 bit) for CMP detection. The primary image information consisting of distinct grey value distributions for every protein can then be described as a relatively simple geometric structure by using combinatorial binary vectors in X/Y/Z (3D) or X/Y (2D) (Figure 43.2b). Each CMP represents the corresponding topologically determined arrangement of proteins (or of other molecular components) in every single compartment of the cell. These vectors can be directly transformed as geometric objects and as a functional map of the whole cell. Such maps are termed toponome maps. Figure 43.2b illustrates schematically CMP lists of two cells. Distinct CMPs can have one (or also several) protein(s) in common (= 1;

so called lead proteins); one or several proteins can be absent (= 0 = anti-colocated); some proteins can be variably existing together with the lead protein (= * = wild card proteins). These common features of different CMPs are called the CMP motif. As revealed by direct comparison of the cells in Figure 43.2b these cells use the same proteins to generate distinct CMP motifs. Following the toponome theory, the cells thereby encipher distinct selectivities of the cell surface that can play important differential functions in cell to cell or cell to matrix interactions, or within cells to encipher a complex directed action, such as driving a tumor cell from a spherical state into an explore state preceding migration. A toponome mapping example of three-dimensional multiprotein complexes on the cell surface of a human blood T cell (3D CMPs) corresponding to the scheme in Figure 43.2 is shown in Figure 43.4e. Similarly, a non-threshold based real-time 100-component biomolecular profiling within an intact tissue section across the human skin reveals the cell surface-coupled 100-dimensional supermolecule linking the cell surface of the basal keratinocytes with the extracellular matrix (Figure 43.4, PPP 3) in a disease (psoriasis) and the non-involved skin (PPP 0) of the same patient. The remarkable issue here is that disease specific supermolecule is directly visualized in context with the normal case (same patient, same genome), and serves as proof of principle for such approaches.

These examples show that toponomics provides direct access to the functional organization of large molecular systems *in situ*.

CMP combinatorial molecular phenotype.

CMP motif Annotation (1; 0; *) describing a characteristic feature of a cluster of distinct CMPs: lead protein (1; = L = a molecule that is present in all CMPs of a cluster); a molecule that is always absent in a motif (0; = A = anti-colocated); molecules present in at least one, but not in all CMPs (*; = W = wild card variably associated with L).

43.1.3 Summary and Outlook

Toponomics is a non-destructive technology essential for direct access to natural large molecular systems in tissues and cells, thereby introducing a "molecular systems histology" and "molecular systems cell biology." The ICM technology for toponome mapping is mature and has been established within the human toponome project in Europe and in the USA (www.huto.toposnomos.com). Cyclical imaging, the principle of ICM-toponomics, has been applied to address various biological problems, and mathematics/informatics approaches to ICM data have been published. Several editorials and a research highlight have featured toponome research. Toponomics reveals the experimentally and clinically validated fact that natural *in situ* molecular systems have their own topological rules that cannot be derived from the isolated parts: the systems structure and function cannot be predicted from their isolated parts, because the relational (emergent) properties possessed by the system as a whole are not possessed by any of its isolated components. The toponomics discovery of disease specific hierarchies within large *in situ* molecular systems controlled by lead protein(s), and the successful therapeutic addressing of such proteins resulting in positive clinical effects, has paved the way to a systematic large-scale toponomics finding of the right lead proteins across major diseases. The driving force is the finding that therapeutic downregulation of a disease specific lead protein not only leads to disassembly of the disease driving molecular robustness network, as shown for tumor cells, but also to a halt of disease progression of many patients as shown for neurologic disease, as known so far. Both experimental and clinical validation of toponomics together with novel highly efficient affinity reagents, such as recombinant human antibodies, recognizing multiple protein domains per protein in thousands of proteins in one toponome mapping experiments is now feasible and will be established as an efficient target finding strategy and *in situ* decoding of disease mechanisms.

PPP Pixel protein profiling. As an extension of threshold based CMPs (1 bit per protein) direct *in situ* protein profiling by using a ICM similarity mapping approach represents the realization of large-scale *in situ* proteomics with the power of combinatorial discrimination (PCMD) per data point of up to $65\,536^k$, where 65 536 is the number of grey value levels (output of a 16-bit CCD camera) and k is the number of co-mapped biomolecules and/or subdomains per biomolecule(s).

Acknowledgements

This research was supported by the Klaus Tschira Foundation (KTS) through projects Toponome Atlas I and II, the Deutsche Forschungsgemeinschaft (DFG Schu627/10-1), the BMBF (grants CELLECT, NBL3, NGFN2, and NGFNplus), the DFG-Innovationskolleg (INK15) as well as the EU project IMAGINT (Health-F5-2011-259881), and the human toponome project (www.huto.toposnomos.com). I thank ToposNomos Ltd. for providing access to the Imaging Cycler® reference laboratory and Andreas Krusche and Reyk Hillert for helping with the figures and formatting the manuscript.

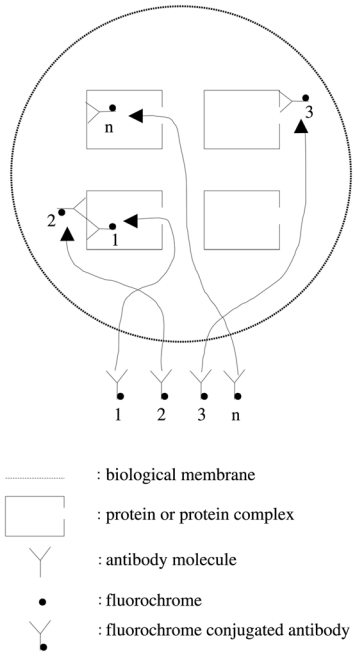


Figure 43.5 Simple scheme of the cellular organization as fundament for toponome mapping by ICM: protein "spaces" (angled symbols) and aqueous "spaces" (free spaces, aqueous compartments) are highly conserved in a cell (principle of city of Venice with houses and water channels). After gentle fixation of this state using appropriate methods, which make it possible at the same time for macromolecules to diffuse across cellular membranes, a quasi-randomly large number of distinct fluorochrome (e.g., FITC) conjugated antibodies (or any specific affinity reagent) can be applied for incubation on the sample sequentially (repetitive incubation–imaging–bleaching cycles), a procedure that is called Venicing in the laboratory jargon. After diffusion across the aqueous compartments the dye-conjugated antibodies recognize and bind to their cognate epitopes. Procedure: (e.g., 1–4 in the figure): Fluorochrome-conjugated antibody is incubated for binding the corresponding epitope, then the corresponding specific fluorescence signals are registered, which is followed by soft-bleaching of the signal at low sunlight – similar natural stray light energy (avoiding lasers): the sunlight hypothesis of ICM. The Venice principle and the sunlight hypothesis are represented as natural biophysical conditions in an ICM to enable large-scale biomolecule *in vivo*/*in situ* profiling without interfering with the tissue and biomolecule structures.

Mass Spectrometry, Chapter 15

43.2 Mass Spectrometry Imaging

Bernhard Spengler

Justus Liebig University at Giessen, Heinrich-Buff-Ring 17, 35392 Giessen, Germany

43.2.1 Analytical Microprobes

Methods in high-throughput proteomics and metabolomics have demonstrated the power and efficiency of fast and accurate detection and identification of compounds in complex samples. The vast amount of data generated with this strategy, however, has also revealed that the bulk molecular composition of a sample is unable to give final and convincing answers to striking questions about cellular processes. Besides qualitative and quantitative analysis, the *localization* of specific compounds and markers in organs, tissues, and cells has been found to be of paramount importance for the understanding of natural states and processes of life. Such compounds, being indicators for a specific biochemical or physiological context, might be inorganic ions (e.g., Ca^{2+}, Fe^{2+}), trace elements (e.g., selenium, cobalt), pharmacologically active ingredients, lipids, oligosaccharides, peptides, proteins, or metabolites of biochemical reactions chains. The spatial localization of such compounds or "targets" can be attributed as "toponomics." The toponome describes the set of spatial concentration distributions of members of a certain class of substances (e.g., all proteins in a biological cell). In a more specific definition, a toponome represents the visualization of an interaction or co-localization of two or more substances in, for example, heterogeneous protein complexes, and the principles of their appearance (Section 43.2.6).

The goal of describing biochemical processes not only qualitatively and quantitatively but also toponomically is not new. Fluorescence microscopy combined with immunochemical labeling, for example, is a well-established technique for visualizing physiological situations. It is understood as an intrinsic disadvantage of such labeling-based imaging techniques that the target molecules have to be known and selected prior to analysis, to allow for defining an appropriate and selective marker (such as an antibody). The laborious and restricting labeling step can be avoided, if the chemical composition of a targeted biological sample region can be determined without specific sample preparation steps and without substance-specific knowledge. This, in contrast to labeling-based imaging techniques, is possible by using mass spectrometry as the source of analytical information. Soon after the introduction of high-power laser systems, the first attempts at spotwise analysis of biological material by mass spectrometric laser-induced desorption of ions were reported in the 1970s, using ultraviolet, pulsed laser radiation. This technique, called laser microprobe mass analysis (LAMMA), is still being used in a number of laboratories, for example, for the detection of trace elements in biological samples.

Similarly, secondary ion mass spectrometry (SIMS) has been able quite early on to detect atoms and small molecules on the surface of technical or biological surfaces. The two methods have a lateral resolution in the range of about 1 µm, but initially were not able to rasterize samples and visualize concentration images, but only provided punctual information about samples. Only with the introduction of the so-called *ion imaging* mode of SIMS did concentration images of elements and small molecules became available.

In the field of mass spectrometric detection of intact biomolecules, the developments of matrix-assisted laser desorption ionization (MALDI) by Franz Hillenkamp and Michael Karas and of electrospray ionization (ESI) by John Fenn in the 1980s directed the focus towards bulk analysis of large molecules, such as proteins or protein complexes. Then, in the 1990s, interest in a direct analytical visualization of native samples returned. The proof-of-principle of MALDI mass spectrometry imaging of biomolecules was first described in 1994 by Spengler *et al*. for high-lateral-resolution imaging of peptide standards. In the following years, numerous applications of this principle were reported (Figure 43.6).

43.2.2 Mass Spectrometric Pixel Images

The generation of microanalytical pixel images of biological samples requires several computer intensive processing steps. The minimal procedure includes the following (Figure 43.7): A highly focused, pulsed laser beam (or primary ion beam, in the case of SIMS) is directed to an exactly positioned sample spot. This results in the formation and desorption of ions, which are

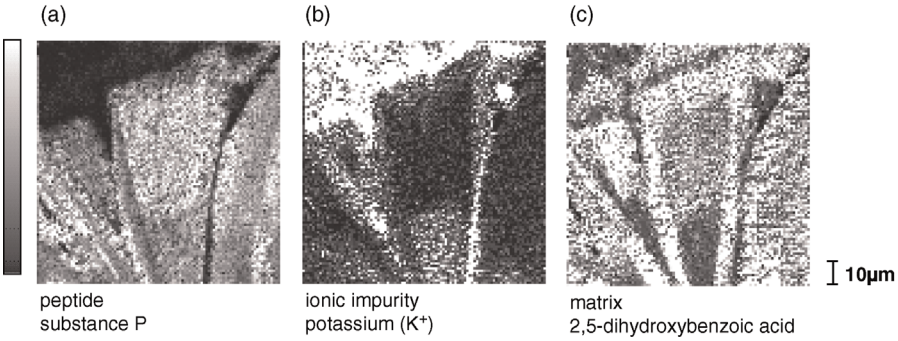


Figure 43.6 Highly resolved visualization of the spatial distribution of a peptide within a standard MALDI preparation (a); distribution of salt impurities (b); distribution of the matrix 2,5-dihydroxybenzoic acid (c). The figure demonstrates the exclusion principle of ion formation, as peptide ions are formed only from micro-areas, which are free from alkali ions. Courtesy of B. Spengler, Giessen.

then mass analyzed and their signals stored as a mass spectrum for this sample spot. Spot diameters are typically in the micrometer range, pulse durations in the nanosecond range. The focus diameter of the laser beam (which results in an ablation spot of a similar size) is typically taken as the step size of the sample manipulation stage and thus as the pixel size of the resulting image. The step sizes can also be larger or smaller than the focus diameter, in order to create an overview image with a lower lateral resolution, or to enhance the achievable lateral resolution by so-called oversampling data acquisition, with step sizes smaller than the ablation spot. The next step after acquisition of a spot-related mass spectrum is to move the sample stage to the next position. In SIMS, it is not the sample that is moved but the ion beam is scanned across the sample surface. When the complete area is scanned and mass analyzed, the resulting two-dimensional data set can be searched for useful mass signals, which represent sample components and which are suitable for visualization as an image. All these signals can now be transformed into gray scale images, with signal intensities coded as gray levels between black and white for each pixel of the image. Color coding can be used to visualize a broader dynamic range, since more colors than gray values can be differentiated by the human eye. Colors can also code different components in one image. Co-localization of compounds, differential display of concentration distributions, or complex correlations can be displayed with colors. Red-green-blue (RGB) images are used in this context to visualize three selected compound distributions in one image without information loss (Figure 43.8).

With all visualization procedures it is especially important for a powerful imaging software to always retain the original data and to easily provide access to the raw data, correlated with the final visual representation.

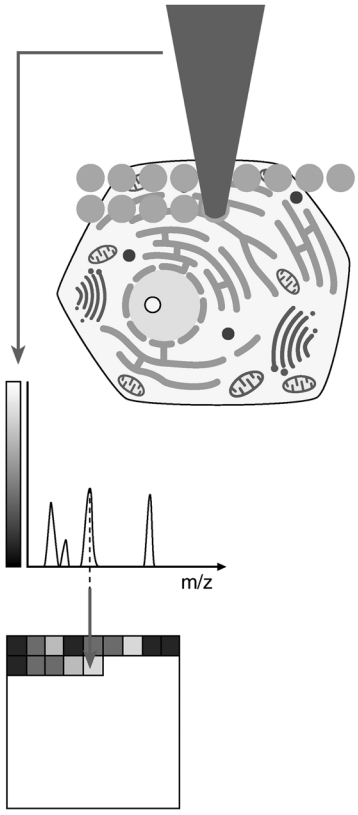


Figure 43.7 Generation of a microanalytical image from position-related mass spectrometric data.

43.2.3 Achievable Spatial Resolution

The spatial resolution that can be achieved with an imaging method is always a key parameter in toponomy. A lateral resolution of 25–50 µm can easily be obtained with MALDI instrumentation, without much technical effort. Reaching a cellular resolution down to 1 µm, however, is hindered by principle limitations, resulting from physical laws of optics and properties of light. To distinguish between low-resolution and high-resolution MALDI imaging techniques, the term "scanning microprobe MALDI (SMALDI)" was introduced for the high-resolution method. Differences between the two are found not only in the analytical goal (regarding the size of structures to be resolved) but especially in the underlying mechanisms of material evaporation and ionization. The principle of usability of MALDI for desorption, ionization, and imaging of biomolecules in the laser focus range of a micrometer was first demonstrated in 1994 (Figure 43.6). As with normal MALDI mass spectrometry of biomolecules it is necessary for SMALDI that matrix and analyte mix in the liquid (solution) phase and that analyte molecules are embedded in the growing matrix crystals upon desolvation of the matrix droplets. (There are, however, certain substance classes of smaller biomolecules, such as phospholipids and carbohydrates, that do not necessarily require this embedding process, even if the limit of detection is improved by it.) Optimization of this embedding process in normal MALDI analyses is of interest primarily with respect to achievable signal intensity (i.e., limit of detection). On the micrometer scale, however, this process has a dramatic influence on the morphology and topology of the sample components. High-resolution images of dried-droplet MALDI samples clearly visualize these phenomena (Figure 43.6). The topologic distribution of

MALDI-MS, Section 15.1.1

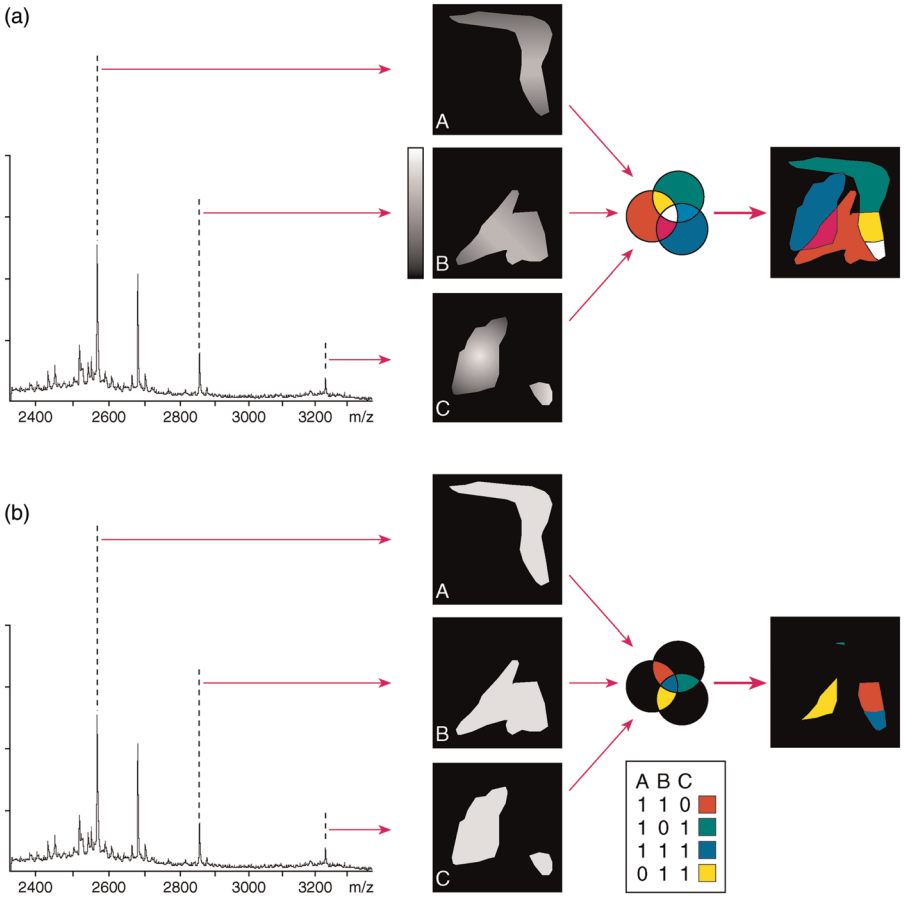


Figure 43.8 Different ways of visual coding of mass spectrometric position information. The spatial distribution of three components A, B, C can first be visualized as gray scale images ((a), middle column). White, in this case, codes high signal intensity and black low intensity. Up to three components can furthermore be coded in parallel in one RGB (red, green, blue) image, without information loss, if each of the three components is coded in one of the three native color channels ((a), right-hand side). The local coincidence of several components can, on the other hand, be visualized by logic operations of signal values of a large number of components and displayed with a large number of chosen colors (b). In this case, not signal intensities of individual components are coded directly but the presence of signal intensities above a certain threshold (b), right). This is a non-lossless information transformation that allows presentation of a visualization of a large number of channels, but with reduced information contents.

the peptide substance P in the dried-droplet MALDI sample reveals its predominant localization within the crystals of the matrix 2,5-dihydroxybenzoic acid (DHB). In contrast, alkali ions, which are present in the sample as impurities, are found outside these crystals and not within them. Highly polar or ionic substances are known to hinder the detection of peptide ions in MALDI-MS in general. Matrix preparation thus can be understood as a micro-separation step that cleans the microregions of the individual matrix crystals from interfering ionic substances. Desorption and ionization of peptides can subsequently take place from alkali-free microcrystal regions under highly focused laser irradiation. In low-resolution MALDI instruments, this effect of mutual exclusion cannot be observed with the same clearness, as the large laser focus always covers multiple matrix crystals and alkali-enrichment areas at the same time. In SMALDI-MS, on the other hand, the mutual exclusion of biomolecular ion signals and ionic impurities signals is obviously very distinct and is observed very clearly.

Inclusion of analyte molecules in matrix crystals is a necessary prerequisite for a sensitive detection of peptides and proteins. It is therefore essential for the detection of such biomolecules from biological tissues or cells by MALDI or SMALDI mass spectrometry to allow for a mixing of matrix and analyte in solution, at least for a short while, and thus to allow doped matrix crystals to grow. This process, however, will necessarily lead to a blur of originally localized analyte concentrations and to the formation of artifacts of concentration peaks of certain analytes, due to segregation effects (Figure 43.9). This situation reveals a dilemma for the goal of highly sensitive and highly laterally resolved analysis of tissue. It can be solved only with a compromise by choosing acceptable conditions in terms of achievable sensitivity and lateral resolution. Such a compromise has been found, leading to an effective lateral resolution of 3 µm at a sufficiently low limit of detection for biological tissue, by spraying or vapor-depositing matrix with dedicated protocols. There is also a search for alternative methods of mass spectrometry imaging that work without application of a dedicated matrix compound. One of these methods is infrared laser desorption/ionization mass spectrometry, which is currently under development.

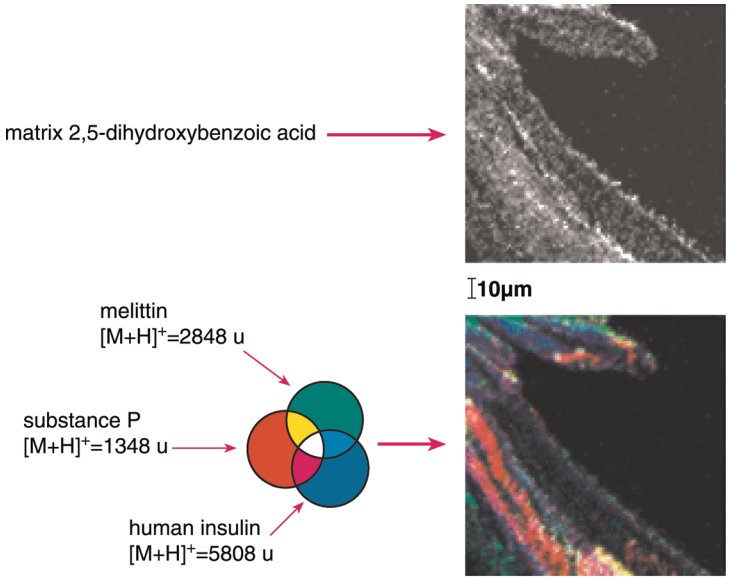


Figure 43.9 SMALDI mass spectrometry imaging of a dried-droplet preparation of a peptide mixture of substance P, melittin, and insulin. The images show that analyte components are dislocated within the sample, as a result of the matrix application. The originally homogeneous solution of peptide components is turned into an inhomogeneous sample after matrix addition and crystallization. Such artifact formation has to be taken into account in any MALDI analysis.

43.2.4 SIMS, ME-SIMS, and Cluster SIMS Imaging: Enhancing the Mass Range

In contrast to MALDI, the detection sensitivity in secondary ion mass spectrometry is a clear function of analyte mass, leading to a loss of detectability of larger biomolecules. Recent developments, however, have demonstrated significant improvements. The application of a matrix to a tissue sample is one possible way to extend the mass range in SIMS, by achieving a MALDI-like desorption/ionization process. This method, called matrix-enhanced SIMS (ME-SIMS), has been used for mass spectrometry imaging, but is limited in spatial resolution, due to the same matrix-related effects as SMALDI imaging.

Another improvement of SIMS was obtained by the development of cluster ion sources, using C_{60} ions or large clusters of gold or argon. These primary ion sources do not reach the same high lateral resolution as classical SIMS sources, but at least allow desorption and ionization of small biomolecules with a lateral resolution in the low micrometer range. Still, SIMS methods are unable (in bulk or imaging mode) to detect intact larger biomolecules such as proteins.

43.2.5 Lateral Resolution and Analytical Limit of Detection

Matrix crystallization is only one of the mechanistic limitations in the achievable lateral resolution. Independent of the individual ionization process (i.e., for MALDI just as for SIMS) there is a fundamental restriction in further improving lateral resolution, which is based on the decreasing number of accessible analyte molecules with decreasing focus area. Assuming a number of six billion peptide molecules under a focus diameter of 200 µm in one monolayer, this number is reduced to only 200 000 molecules in a 1 µm spot. Depending on instrument type, a mass spectrometer requires between 1000 and 100 000 analyte molecules in the sample for generation of a useful mass spectrum. The limitation is to some extent eased by the fact that in mass spectrometry not only the absolute signal intensity defines the quality of a mass spectrum but even more so the signal-to-noise ratio. Fortunately, the intensity of noise signals (especially those of chemical noise) also decreases with decreasing sampling area. Nevertheless, there still is a final limit of detection that cannot be overcome, namely, due to signal statistics at too low molecule numbers. One also has to take into account that, typically, not complete monolayers of only one analyte component are to be analyzed but complex mixtures of a large number of different components. Low abundance molecules in biological tissue or on cell surfaces are thus naturally problematic for highly resolved mass spectrometry imaging. A useful lateral resolution significantly below 1 µm will therefore be achievable by MALDI or SIMS imaging only for abundant biomolecules due to fundamental reasons.

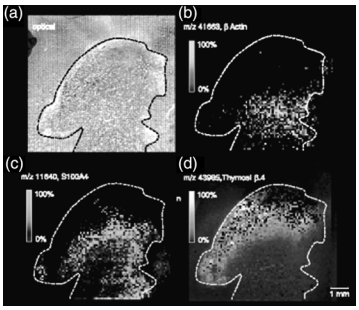


Figure 43.10 MALDI images of a glioblastoma tissue section, covered with sinapinic acid as a matrix and scanned with a step size of 100 µm. The images indicate an accumulation of thymosin-β.4 in the proliferating area of the tumor (d), while other proteins were mostly found in the ischemic and necrotic areas ((b) and (c)). Courtesy by R. Caprioli, Vanderbilt University.

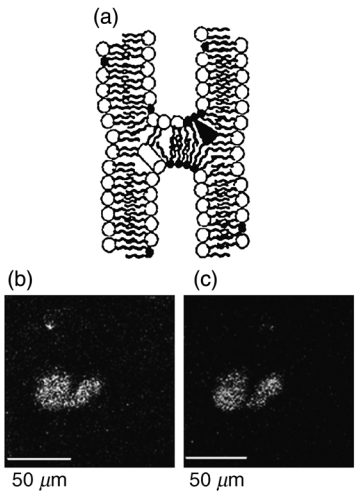


Figure 43.11 SIMS images of two protozoa during mating (conjugation). A decrease of phosphatidyl choline was found in the conjugation zone (c), while the total concentration of phospholipids was found to stay constant (b). The scheme (a) indicates the assumed distribution of phosphatidyl cholines (white circles) and phosphatidyl ethanolamines (black circles). Courtesy by A.G. Ewing, Pennsylvania University.

43.2.6 Coarse Screening by MS Imaging

Mass spectrometry imaging currently is still in a state of methodological evolution. Early-published applications looked promising but mostly did not include proofs of validity, reproducibility, and biomedical information of analytical images. Mass spectrometric data revealed molecular topologies within tissue images, but the identity and homogeneity of underlying substance signals could not be unveiled. Owing to these methodological and technological limitations, MALDI imaging and SIMS imaging initially remained purely qualitative and descriptive methods, unable to meet the conditions of analytical-chemistry standards. Results were reported using MALDI in the low lateral-resolution range and using SIMS in the low mass range.

Protein imaging on a qualitative and unvalidated level were reported at first, with a lateral resolution of about 100 µm. One early example describes the analytical imaging of a mouse brain (Figure 43.10). Variations in the lateral distributions were found for various expected proteins in the different brain regions. In the vicinity of a glioblastoma, a significant increase of a signal attributed to thymosin-β-4 was found. Identification of the suspected signal could not be performed based on the mass spectrometric imaging data, but was obtained by classical LC-MS/MS analysis of trypsin-digested tumor tissue in a parallel study. The identity of proteins determined in LC-MS/MS runs with those of according *m/z* signals in imaging experiments was tacitly assumed.

Bioanalytical applications at a high lateral resolution in the micrometer or sub-micrometer range were mostly limited to SIMS methodology, in a rather low mass range. The high spatial resolution of secondary ion mass spectrometry allows us to visualize processes in the cell membrane, provided that investigated analytes are detectable with sufficient sensitivity. A remarkable example is the visualization of the mating of protozoa *Tetrahymena thermophila*, which reveals the conjugation of cell membranes for the exchange of genetic material. The zone of conjugation showed a significant reduction of phosphatidyl choline concentrations in relation to the total concentration of phospholipids (Figure 43.11). This indicates a severe reordering of lipids in the hyphenation zone. In this investigation, again, the identification of signals as being phosphatidyl choline fragments had to be performed by separate tandem mass spectrometric analysis. It was not possible to distinguish and characterize the imaged phosphatidyl choline species directly in the membranes, since only the head group of the phosphatidyl cholines was detected as a fragment and not the intact molecules with their isomeric fatty acids.

43.2.7 Accurate MALDI Mass Spectrometry Imaging

Not only with SIMS, but also with SMALDI mass spectrometry imaging, cellular and sub-cellular resolution is achievable. In contrast to the SIMS method, the mass range is not a prominent limitation of biomolecular analyses in SMALDI imaging. Cell cultures of a human renal cell carcinoma have been analyzed with a lateral pixel resolution of 1 µm, showing mass signals up to 15 000 u. Individual cells revealed a reasonable topological structure of analyte signals.

A significant leap in quality of mass spectrometry imaging was obtained by the recent combination of high lateral resolution, high mass resolution and accuracy, and high imaging selectivity on-tissue structure analysis by MS/MS imaging and semi-physiological tissue analysis by atmospheric pressure SMALDI. With that, mass spectrometry imaging was improved and extended to a technique of high validity and high specificity, providing for molecular histology at a high informational level. SMALDI imaging on high resolution mass spectrometers now allows us to distinguish precisely tissue subtypes at a lateral resolution of 2–3 µm. Label-free, non-targeted MALDI mass spectrometry imaging generates hundreds of characteristic tissue signals, extending the performance of bioanalytical tissue analysis considerably, significantly beyond that of histochemical, targeted (label-dependent) methods. Figure 43.12 shows an example of SMALDI imaging of mouse urinary bladder tissue, visualizing three selected tissue-specific signals, which obviously enhance the information obtained from histological staining.

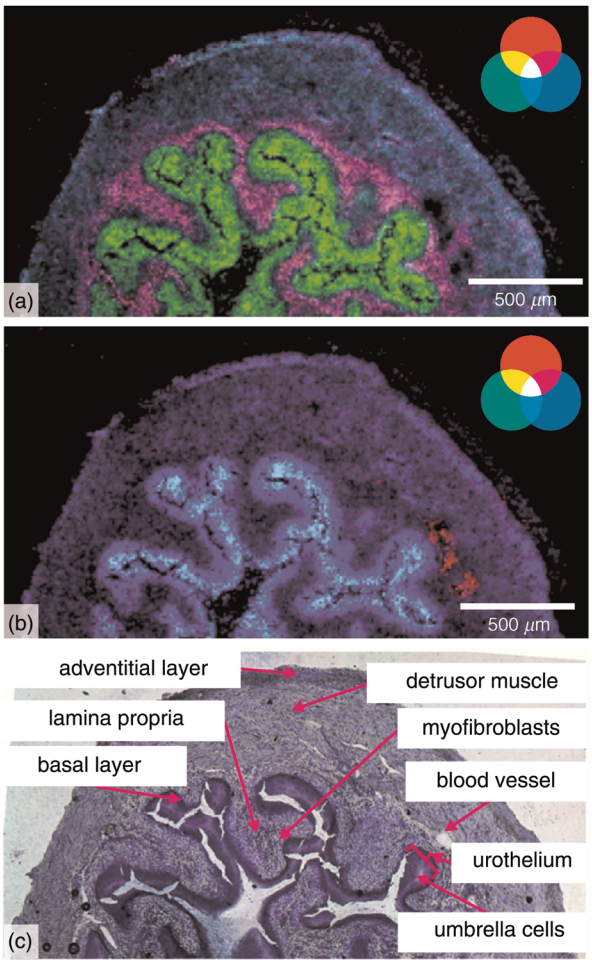


Figure 43.12 SMALDI images of a mouse urinary bladder tissue section. (a) 10 μm step size, overlay of ion images of $m/z = 741.5307$ (blue, muscle tissue, sphingomyelin SM(34:1)), $m/z = 798.5410$ (green, urothelium, phosphatidyl choline PC(34:1)), and $m/z = 743.5482$ (red, lamina propria). (b) Overlay of ion images of $m/z = 798.5410$ (blue, urothelium, PC(34:1)), $m/z = 812.5566$ (green, umbrella cells, phosphatidyl ethanolamine PE(38:1)), and $m/z = 616.1767$ (red, blood vessels, heme b, $M^{+\cdot}$). (c) Microscope image of the tissue after staining with toluidine blue. Copied with permission from Römpp, A., Guenther, S., Schober, Y., Schulz, O., Takats, Z., Kummer, W., and Spengler, B. (2010) *Angew. Chem. Int. Ed.*, 3834–3838. Copyright Wiley-VCH Verlag GmbH.

In the field of peptide imaging at a lateral resolution of 5 μm, experiments were reported for neuropeptides in mouse pituitary gland tissue, differentiating individual peptide-containing cells and identifying peptide sequences and post-translational modifications by MS/MS imaging.

43.2.8 Identification and Characterization of Analytes

Mass Analysers, Section 15.2

Mass spectrometry images were limited initially in their information contents to the visualization of an intensity distribution within a certain *m/z* window. Identification and characterization of the imaged substances directly from the image data was not possible. Furthermore, the analytical homogeneity (i.e., the substance-specific unambiguousness of the chosen mass window) was not guaranteed with the employed low-resolution mass analyzers. It was thus rarely possible to perform an analytical validation of the method, prior to any biomedical validation. Instead, in parallel to an imaging analysis, classical identification methods were employed using homogenized material, with a subsequent correlation to the obtained imaging data. This limitation has been eliminated or at least strongly reduced with modern methods of high-resolution and high-accuracy mass spectrometry directly from tissue, allowing for a direct validation of imaging data (Figure 43.13). Coupling imaging ion sources to high-resolution (FT-ICR or orbital trapping) mass spectrometers now provides for a direct identification of individual tissue components, using mass-based classification, combinatorial data evaluation and MS/MS imaging.

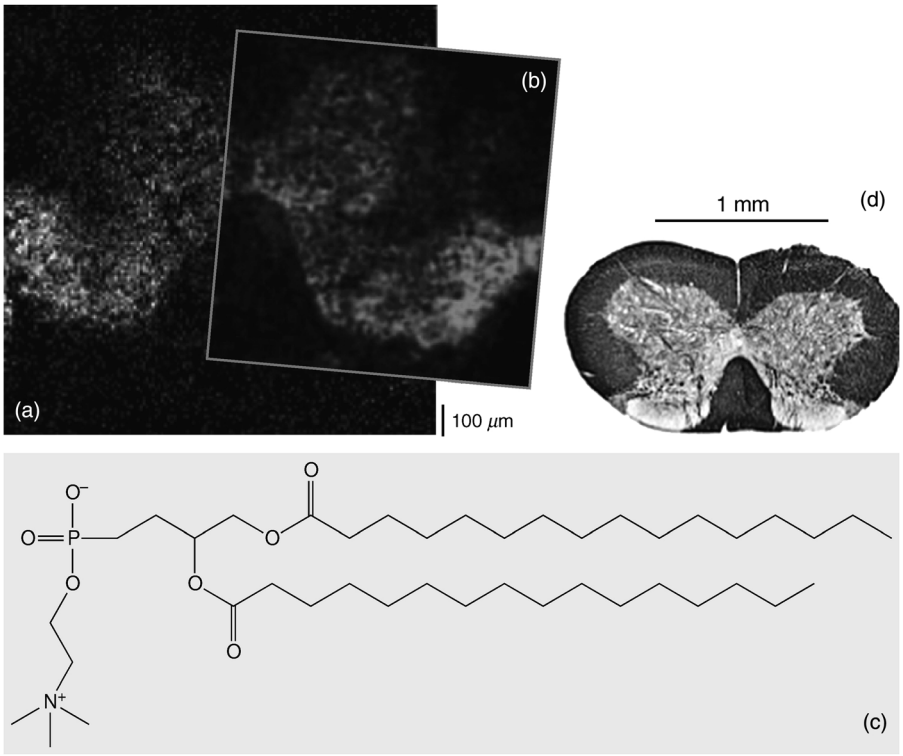


Figure 43.13 SMALDI MS/MS image of mouse spinal cord: (a) 10 μm step size, Orbitrap ($R = 30\,000$), positive ions, MS ion image of $m/z = 772.5253$ (phosphatidyl choline PC(32:0). (b) MS/MS ion image of $m/z = 713.4518$ (precursor ion 772.5253; PC(32:0)). (c) Molecular structure of PC(16:0, 16:0). (d) Microscope image of the tissue after immunostaining. Courtesy by B. Spengler.

Further Reading

Section 43.1

Bode M, Irmler M, Friedenberger M, May C, Jung K, Stephan C, Meyer HE, Lach C, Hillert R, Krusche A, Beckers J, Marcus K, Schubert W. Interlocking transcriptomics, proteomics and toponomics technologies for brain tissue analysis in murine hippocampus. *Proteomics* 2008; **8** (6): 1170–1178. doi: 10.1002/pmic.200700742.

Dress AWM, Lokot T, Pustyl'nikov LD, Schubert W, Poisson Numbers and Poisson Distributions in Subset Surprisology. *Ann Comb*. 2005; **8** (4): 473–485. doi: 10.1007/s00026-004-0234-2.

Dress AWM, Lokot T, Schubert W, Serocka P, Two Theorems about Similarity Maps. *Ann Comb*. 2008; **12** (3): 279–290. doi: 10.1007/s00026-008-0351-4.

Friedenberger M, Bode M, Krusche A, Schubert W. Fluorescence detection of protein clusters in individual cells and tissue sections by using toponome imaging system: sample preparation and measuring procedures. *Nat Protoc*. 2007; **2** (9): 2285–2294. doi: 10.1038/nprot.2007.320.

Hillert R, Gieseler A, Krusche A, Humme D, Röwert-Huber HJ, Sterry W, Walden P, Schubert W. Large molecular systems landscape uncovers T cell trapping in human skin cancer. *Sci Rep*. 2016, **6**: 19012 doi: 10.1038/srep19012.

Nattkemper TW, Ritter HJ, Schubert W. A neural classifier enabling high-throughput topological analysis of lymphocytes in tissue sections. *IEEE Trans Inf Technol Biomed*. 2001; **5** (2): 138–149. doi: 10.1109/4233.924804.

Ostalecki C, Lee JH, Dindorf J, Collenburg L, Schierer S, Simon B, Schliep S, Kremmer E, Schuler G, Baur AS. Multiepitope tissue analysis reveals SPPL3-mediated ADAM10 activation as a key step in the transformation of melanocytes. *Sci Signal*. 2017, **10** (470). pii: eaai8288. doi: 10.1126/scisignal.aai8288.

Schubert W. Multiple antigen-mapping microscopy of human tissue. In: Burger G, Oberholzer M, Vooijs GP editors. Excerpta medica. *Adv Anal Cell Pathol*. Elsevier, Amsterdam; 1990. p. 97–98.

Schubert W. Topological Proteomics, Toponomics, MELK-Technology. *Adv Biochem Eng Biotechnol* 2003; **83**: 189–209. doi: 10.1007/3-540-36459-5_8.

Schubert W, Bonnekoh B, Pommer AJ, Philipsen L, Böckelmann R, Malykh Y, Gollnick H, Friedenberger M, Bode M, Dress AWM. Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nat Biotechnol*. 2006; **24** (10): 1270–1278. doi: 10.1038/nbt1250.

Schubert W. A three-symbol code for organized proteomes based on cyclical imaging of protein locations. *Cytometry A* 2007; **71A** (6): 352–360. doi: 10.1002/cyto.a.20281.

Schubert W, Gieseler A, Krusche A. Hillert R. Toponome Mapping in Prostate Cancer: Detection of 2000 Cell Surface Protein Clusters in a Single Tissue Section and Cell Type Specific Annotation by Using a Three Symbol Code. *J Proteome Res*. 2009; **8** (6): 2696–2707. doi: 10.1021/pr800944f.

Schubert W. On the origin of cell functions encoded in the toponome. *J Biotechnol*. 2010; **149** (4): 252–259. doi: 10.1016/j.jbiotec.2010.03.009.

Schubert W, Gieseler A, Krusche A, Serocka P, Hillert R, Next-generation biomarkers based on 100-parameter functional super-resolution microscopy TIS. *N Biotechnol*. 2012; **29** (5): 599–610. doi: 10.1016/j.nbt.2011.12.004.

Schubert W. Toponomics. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H, editors. *Encyclopedia of Systems Biology*. New York: Springer; 2013. p. 2191–2212. doi: 10.1007/978-1-4419-9863-7_631. ISBN 978-1-4419-9862-0.

Schubert W. Systematic, spatial imaging of large multimolecular assemblies and the emerging principles of supramolecular order in biological systems. *JMR*. 2014; **27** (1): 3–18. doi: 10.1002/jmr.2326.

Schubert W. Advances in toponomics drug discovery: Imaging cycler microscopy (ICM) correctly predicts a therapy method of amyotrophic lateral sclerosis (ALS). *Cytometry Part A;* 2015, **87A**: 696–703. doi: Q6 10.1002/cyto.a.22671.

Section 43.2

Bouschen, W., Schulz, O., Eikel, O., and Spengler, B. (2010) Matrix vapour deposition/recrystallization and dedicated spray preparation for high-resolution scanning microprobe MALDI imaging mass spectrometry (SMALDI-MS) of tissue and single cells. *Rapid Comm. Mass Spectrom.*, **24**, 355–364.

Guenther, S., Römpp, A., Kummer, W., and Spengler, B. (2011) AP-MALDI imaging of neuropeptides in mouse pituitary gland with 5 µm spatial resolution and high mass accuracy. *Int. J. Mass Spectrom.*, **305**, 228–237.

Hillenkamp, F., Unsöld, E.E., Kaufmann, R., and Nitsche, R. (1975) Laser microprobe mass analysis of organic materials. *Nature*, **256**, 119–120.

Hummon, A.B., Sweedler, J.V., and Corbin, R.W. (2003) Discovering new neuropeptides using single-cell mass spectrometry. *Trends Anal. Chem.*, **22**, 515–521.

Jespersen, S., Chaurand, E., van Strien, F.J.C., Spengler, B., and van der Greef, J. (1999) Direct sequencing of neuropeptides in biological tissue by MALDI-PSD. *Mass Spectrom.*, **71**, 660–666.

Luxembourg, S.L., McDonnell, L.A., Duursma, M.C., Guo, X., and Heeren, R.M.A. (2003) Effect of local matrix crystal variations in matrix-assisted ionization techniques for mass spectrometry. *Anal. Chem.*, **75**, 2333–2341.

Ostrowski, S.G., Van Bell, C.T., Winograd, N., and Ewing, A.G. (2004) Mass spectrometric imaging of highly curved membranes during *Tetrahymena* mating. *Science*, **305**, 71–73.

Schröder, W.H. and Fain, G.L. (1984) Light-dependent calcium release from photoreceptors measured by laser micro-mass analysis. *Nature*, **309**, 268–270.

Spengler, B., Hubert, M., and Kaufmann, R. (1994) MALDI ion imaging and biological ion imaging with a new scanning UV-laser microprobe, in Proceedings of the 42nd ASMS Conference on Mass Spectrometry and Allied Topics, Chicago, Illinois, 29 May to 3 June 1994, S. 1041, American Society for Mass Spectrometry

Spengler, B. and Hubert, M. (2002) Scanning microprobe matrix-assisted laser desorption ionization (SMALDI) mass spectrometry: instrumentation for sub-micrometer resolved LDI and MALDI surface analysis. *J. Am Soc. Mass Spectrom.*, 13, 735–748.

Stoeckli, M., Chaurand, E., Hallahan, D.E., and Caprioli, R.M. (2001) Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissue. *Nat. Med.* **7**, 493–496.

Yanagisawa, K., Shyr, Y., Xu, B.J., Massion, E.E., Larsen, E.H., White, B.C., Roberts, J.R., Edgerton, E., Gonzalez, A., Nadaf, S., Moore, J.H., Caprioli, R.M., and Carbone, D.E. (2003) Proteomic patterns of tumour subsets in non-small-cell lung cancer. *The Lancet*, **362**, 433–439.

Appendix 1: Amino Acids and Posttranslational Modifications

Table 1 Amino acids: mass, pK-values and pI values (mass values given are unmodified amino acid mass minus water)

Symbols		Name and Formula	Mono-isotopic mass	Average mass	pK	pI
Ala	A	alanine C_3H_5NO	71.03711	71.07794	2.35 9.69	6.02
Arg	R	arginine $C_6H_{12}N_4O$	156.10111	156.18584	2.17 9.04 (α-amino) 12.48 (guanidino)	10.76
Asn	N	asparagine $C_4H_6N_2O_2$	114.04293	114.10272	2.02 8.8	5.41
Asp	D	aspartic acid $C_4H_5NO_3$	115.02694	115.08744	2.09 (α-carboxy) 3.86 (β-carboxy) 9.82	2.87
Cys	C	cysteine C_3H_5NOS	103.00919	103.14394	1.71 8.33 (Sulfhydryl) 10.78 (α-Amino)	5.02
		cystine $C_6H_{10}N_2O_3S_2$	222.01328	222.28728	1.65 2.26 (Carboxy) 7.85 9.85 (Amino)	5.06
Gln	Q	glutamine $C_5H_8N_2O_2$	128.05858	128.12930	2.17 9.13	5.65
Glu	E	glutamic acid $C_5H_7NO_3$	129.04259	129.11402	2.19 (α-carboxy) 4.25 (β-carboxy) 9.67	3.22
Gly	G	glycine C_2H_3NO	57.02146	57.05136	2.34 9.6	5.97
His	H	histidine $C_6H_7N_3O$	137.05891	137.13940	1.82 6.0 (imidazol) 9.17	7.58
Ile	I	isoleucine $C_6H_{11}NO$	113.08406	113.15768	2.36 9.68	6.02
Leu	L	leucin $C_6H_{11}NO$	113.08406	113.15768	2.36 9.60	5.98
Lys	K	lysine $C_6H_{12}N_2O$	128.09496	128.17236	2.18 8.95 (α-amino) 10.53 (ε-amino)	9.74
Met	M	methionine C_5H_9NOS	131.04048	131.19710	2.28 9.21	5.75
Phe	F	phenylalanine C_9H_9NO	147.06841	147.17390	1.83 9.13	5.98
Pro	P	proline C_5H_7NO	97.05276	97.11522	1.99 10.60	6.10
Ser	S	serine $C_3H_5NO_2$	87.03203	87.07734	2.21 9.15	5.68
Thr	T	threonine $C_4H_7NO_2$	101.04768	101.10392	2.63 10.43	6.53
Trp	W	tryptophan $C_{11}H_{10}N_2O$	186.07931	186.20998	2.38 9.39	5.88
Tyr	Y	tyrosine $C_9H_9NO_2$	163.06333	163.17330	2.20 9.11 (α-amino) 10.07 (phenol)	5.65
Val	V	valine C_5H_9NO	99.06841	99.13110	2.32 9.62	5.97
Sec	U	selenocysteine C_3H_5NOSe	150.95363	150.03794	5.3 (selenol)	

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

Table 2 Posttranslational modifications and mass changes of modified peptide

Modification	Mono isotopic mass	Average mass
5′-Adenylation	329.0525	329.2091
Acetylation	42.0106	42.0373
N-Acetylhexosamine (GalNAc, GlcNAc)	203.0794	203.1950
N-Acetylneuraminic acid (sialic acid, NeuAc, NANA, SA)	291.0954	291.2579
ADP-Ribosylation (from NAD)	541.0611	541.3052
Biotinylation (via amide bond to Lys)	226.0776	226.2994
Carboxylation (of Asp and Glu)	43.9898	44.0098
C-terminal amide (from Gly)	−0.9840	−0.9847
Cysteinylation	119.0041	119.1442
Deamidation (of Asn and Gln)	0.9840	0.9847
Deoxyhexoses (Fuc, Rha)	146.0579	146.1430
Disulfide bridge	−2.0157	−2.0159
Farnesylation	204.1878	204.3556
Formylation	27.9949	28.0104
Geranylgeranylation	272.2504	272.4741
Glutathionylation	305.0682	305.3117
N-Glycolylneuraminic acid (NeuGc)	307.0903	307.2573
Hexosamines (GalN; GlcN)	161.0688	161.1577
Hexoses (Fru, Gal, Glc, Man)	162.0528	162.1424
Homoserine (from Met, by CNBr treatment)	−29.9928	−30.0935
Hydroxylation	15.9949	15.9994
Lipoic acid (via amide bond to Lys)	188.0330	188.3147
Methylation	14.0157	14.0269
Myristoylation	210.1984	210.3598
Oxidation (of Met)	15.9949	15.9994
Palmitoylation	238.2297	238.4136
Pentoses (Ara, Rib, Xyl)	132.0423	132.1161
4-Phosphopantetheine	339.0780	339.3294
Phosphorylation	79.9663	79.9799
Proteolysis (of a peptide bond)	18.0106	18.0153
Pyridoxal phosphate (via Schiff base to Lys)	231.0297	231.1449
Pyroglutamic acid (from Gln)	−17.0265	−17.0306
Stearoylation	266.2610	266.4674
Sulfation	79.9568	80.0642

Appendix 2: Symbols and Abbreviations

2DPP-mapping	two-dimensional phosphopeptide mapping
4-NA	4-nitroanilide
^{5ca}C	5-carboxylcytosine
^{5f}C	5-formylcytosine
^{5hm}C	5-hydroxymethylcytosine
$5^{m}C$	5-methylcytosine
^{5m}CG	CpG-methylation
α	selectivity
ACE	affinity capillary electrophoresis
ACQ	6-aminoquinoyl-*N*-hydroxysuccinimidyl carbamate
AD	activation domain
AFM	atomic force microscopy
AFM	atomic force microscope
ALS	amyotrophic lateral sclerosis
AMV	avian myeloblastosis virus
AO	atomic orbital
AOBS	acousto-optical beam-splitter
AOTF	acousto-optic tunable filter
AP	alkaline phosphatase
AP-MS	affinity purification mass spectrometry
APSY	automated projection spectroscopy
ARM	arginine-rich motif
ASKA	analogue-sensitive kinase allele
ATH	amino acid thiohydantoin
ATP	adenosine triphosphate
ATR	attenuated total reflection
ATZ	anilinothiazolinone
β_2m	β_2-microglobulin
BACTH	bacterial adenylate cyclase-based two-hybrid system
BCA	bicinchoninic acid
BCIP	5-bromo-4-chloro-indoxyl phosphate
bDNA	branched DNA amplification
BFP	blue fluorescent protein
BIO	biotin:avidin (or streptavidin)
BIOS	biology-oriented synthesis
BLAST	Basic Local Alignment Search Tool
BNPS-skatole	3-bromo-3-methyl-2-(2-nitrophenylthio)-3*H*-indole
BOD5	biochemical oxygen demand in 5 days
BSA	bovine serum albumin
BSTFA	bis(trimethylsilyl)trifluoroacetamide
BTK	Bruton tyrosine kinase
$C^{5m}CWGG$	Dcm-methylation
CAD	collision-activated dissociation
CAP	catabolite activator protein
CAT	chloramphenicol acetyltransferase
CBP	calmodulin-binding peptide
CCD	charge-coupled device
CD	circular dichroism
CD	cyclodextrin
CDR	complementarity-determining region
CE	capillary electrophoresis
CE	collision energy
CEC	capillary electrokinetic chromatography
CEC	capillary electrochromatography
CEkF	capillary electrokinetic fractionation
CEM	channel electron multiplier
CET	cryo-electron tomography
CF	cystic fibrosis
CFP	cyan fluorescent protein
CGE	capillary gel electrophoresis
CGH	comparative genomic hybridization
CHEF	contour-clamped homogeneous electric field
ChIP	chromatin immunoprecipitation
ChIP chip	chromatin immunoprecipitation and chip analysis
CI	chemical ionization
CID	collision-induced dissociation
CIEF	capillary isoelectric focusing
CISS	chromosomal *in situ* suppression
cLSM	confocal laser scanning microscopy
CMC	critical micelle concentration
CMP	combinatorial molecular phenotype
CNV	copy number variant
CoBRA	combined bisulfite restriction analysis
co-IP	co-immunoprecipitation
COSY	correlation *s*pectroscop*y*
CRISPR	clustered regulatory interspaced short palindromic repeats

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

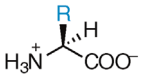
CRM	charged-residue model
CSP	cold-shock protein
CTAB	cetyltrimethylammonium bromide
CTF	contrast transfer function
CZE	capillary zone electrophoresis
DABS-Cl	4-dimethylaminoazobenzene-4-sulfonyl chloride
DBD	DNA-binding domain
DBE	direct blotting electrophoresis
DDA	data-dependent acquisition
DECP	diethyl pyrocarbonate
DFP	diisopropyl fluorophosphate
DFT	density functional theory
DGGE	denaturing gradient gel electrophoresis
DHR	dihydro-rhodamine
DIA	data-independent acquisition
DIC	*N,N′*-diisopropylcarbodiimide
DIG	digoxigenin:anti-digoxigenin
DIGE	difference gel electrophoresis
DMPC	dimyristoyl-phosphatidylcholine
DMS	dimethyl sulfate
DMT	dimethoxytrityl
DNMT	DNA methyltransferase
DNP	2,4-dinitrophenol
DO	dissolved oxygen
DOS	diversity-oriented synthesis
DPTU	diphenylthiourea
DR	dummy residues
ds	double-stranded
DSC	differential scanning calorimeter
DSC	differential scanning calorimetry
dsRBD	double-stranded RNA binding domain
EBV	Epstein–Barr virus
ECD	electron capture dissociation
EDS	energy-dispersive X-ray spectrometer
EDTA	ethylenediaminetetraacetic acid
EELS	electron energy loss spectroscopy
EGFR	epidermal growth factor receptor
EI	electron impact
EI	electron impact ionization
EIA	enzyme-immune assay
EIC	extracted ion chromatogram
ELISA	enzyme-linked immunosorbent assay
ELSD	evaporative light scattering detector
EMBL	European Molecular Biology Laboratory
EMBOSS	European Molecular Biology Laboratory Open Software Suite
EMSA	electrophoretic mobility shift analysis
ENDOR	electron nuclear double resonance
ENU	ethylnitrosourea
EOF	electroosmotic flow
EOF	endoosmotic flow
EOM	ensemble optimization method
EPL	expressed protein ligation
EPO	erythropoietin
EPR	electron paramagnetic resonance
ER	endoplasmic reticulum
ESEEM	electron spin echo envelope modulation
ESI	electrospray ionization
ESI	electron spectroscopic imaging
ESI-MS	electrospray ionization mass spectrometry
ESR	electron spin resonance
EST	expressed sequence-tagged
ETD	dissociation electron transfer
ETD	electron transfer dissociation
ETR	electron transfer reaction
F	flow rate
FAB	fast atom bombardment
FACS	fluorescence-activated cell sorting
FCM	fluorescence correlation microscopy
FDA	Food and Drug Administration
FDR	framework-determining region
Fe-BABE	Fe 1-(*p*-bromoacetamidobenzyl)ethylene-diaminetetraacetic acid
FEL	free electron laser
FET	field effect transistor
FFT	fast Fourier transform
FIB	focused ion beam
FID	free induction decay
FID	flame ionization detector
FIGE	field inversion gel electrophoresis
FIR	far-infrared
FISH	fluorescence *in situ* hybridization
FITC	fluorescein isothiocyanate
FLIM	fluorescence lifetime imaging microscopy
FLUGS	fluorescein:anti-fluorescein
FMLP	*N*-formyl-Met-Leu-Phe
FMOC	fluorenylmethoxycarbonyl
FRAP	fluorescence recovery after photobleaching
FRET	Förster resonance energy transfer
FRET	fluorescence resonance energy transfer
FSM	fluorescent speckle microscopy
FT	Fourier transform
FT-ICR-MS	Fourier-transform ion cyclotron resonance mass spectrometer
FT-IR	Fourier transform infrared
FT-NMR	Fourier-transform NMR
FWHM	full-width at half-maximum
GCC	graphitized carbon column
GFP	green fluorescent protein
$G^{N6m}ATC$	Dam-methylation
GOD	glucose oxidase
gRNA	guide RNA
GSD	ground-state depletion
GST	glutathione-*S*-transferase
GWAS	genome-wide association studies
HAT	acetyl transferase
HBsAg	hepatitis B surface antigen
HCD	higher-energy collisional dissociation
HCV	hepatitis C virus
HDAC	histone deacetylase
HDR	homology-directed repair
HETP	height equivalent to one theoretical plate
HGP	Human Genome Project
HIC	hydrophobic interaction chromatography
HIV	human immunodeficiency virus
HLH	helix-loop-helix protein
HOBt	1-hydroxybenzotriazole
HP-AC	high-performance affinity chromatography
HP-AEX	high-performance anion exchange chromatography

HPAEC-PAD	high performance anion exchange chromatography with pulsed amperometric detection
HP-ANPC	high-performance aqueous normal-phase chromatography
HP-CEX	high-performance cation exchange chromatography
HP-GPC	high-performance gel permeation chromatography
HP-HIC	high-performance hydrophobic interaction chromatography
HP-HILIC	high-performance hydrophilic interaction chromatography
HP-IEX	high-performance ion exchange chromatography
HPLC	high-performance liquid chromatography
HP-NPC	high-performance normal-phase chromatography
HP-RPC	high-performance reversed-phase chromatography
HP-SEC	high-performance size exclusion chromatography
HPTLC	high-performance thin-layer chromatography
HSQC	heteronuclear single-quantum coherence
HTH	helix-turn-helix structure
HVTEM	high-voltage transmission electron microscope
ICAT	isotope-coded affinity tag
ICM	imaging cycler microscopy
ICPL	isotope-coded protein label
IDP	intrinsically disordered protein
IEC	ion exchange chromatography
IEF	isoelectric focusing
IEM	ion evaporation model
IEX	ion exchange chromatography
IMAC	immobilized metal chelate chromatography
IMS	ion mobility spectrometry
IP	isoelectric point
IP	immunoprecipitation
IPG	immobilized pH gradient
IRE	internal reflection element
IRMPD	infrared multiphoton dissociation
IRS	interspersed repetitive sequence
ISD	in source decay
ISH	*in situ* hybridization
ITC	isothermal titration calorimeter
ITC	isothermal titration calorimetry
ITP	isotachophoresis
iTRAQ	isobaric tags for relative and absolute quantitation
k	retention factor
K	specific conductivity
KLH	keyhole limpet hemocyanin
KTS	Klaus Tschira Foundation
LADH	liver alcohol dehydrogenase
LAMMA	laser microprobe mass analysis
LC	liquid chromatography
LCAO	linear combinations of atomic orbitals
LC-MS/MS	liquid chromatography tandem mass spectrometry
LCR	ligase chain reaction
LCR	locus control region
LD	linear dichroism
LDA	low-density amorphous ice
LDA	linear discriminant analysis
LDI	laser desorption/ionization
LDI-MS	laser desorption ionization mass spectrometry
LED	light-emitting diode
LIF	laser-induced fluorescence
LIMS	laboratory information management system
LINC	light-induced co-clustering
LM	beam path in the light microscope
LNA	locked nucleic acid
LOC	lab-on-a-chip
LOD	lactate oxidase
LSM	laser scanning microscope
LTR	long-terminal repeat
LVSEM	low-voltage scanning electron microscope
μ_{eff}	effective mobility
MALDI	matrix-assisted laser desorption ionization
MALDI TOF MS	matrix-assisted laser desorption/ionization time-of-flight mass spectrometry
MALDI-TOF	matrix-assisted laser desorption/ionization time of flight
MBD	combined bisulfite restriction analysis
MCE	microchip electrophoresis
MCP	microchannel plate
MD	molecular dynamics
MD-HPLC	multidimensional (multistage, multicolumn) high-performance liquid chromatography
MeCP	methyl-CpG-binding protein
MeDIP	methylated DNA immunoprecipitation
MEF	mouse embryonal fibroblast
MEKC	micellar electrokinetic chromatography
MELC	multi-epitope-ligand cartography
MES	minimal ensemble search
ME-SIMS	matrix-enhanced SIMS
MFM	magnetic force microscope
MHC	major histocompatibility complex
MIP	molecular imprinted polymer
MIR	mid-infrared
MIR	multiple isomorphous replacement
MIRA	methylated-CpG island recovery assay
miRNA	microRNA
MMD	multimolecular domain
MMLV Env	Moloney murine leukemia virus envelope
MO	molecular orbital
MOPS	3-*N*-morpholino-1-propane sulfonic acid
MPE-Fe	methidiumpropyl-EDTA-Fe
MPS	massive parallel sequencing
MR	molecular replacement
MRM	multiple reaction monitoring
MS	mass spectrometry
MSA	multistage activation
MSP	methylation-specific PCR
MTF	modulation transfer function

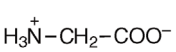
N	plate number
^{N6m}A	N^6-methyladenine
N_A	Avogadro's constant
NA	numeric aperture
NAD^+	nicotine adenine dinucleotide
NAGK	*N*-acetyl-L-glutamate kinase
NAR	nucleic acids research
NASBA	nucleic acid sequence-based amplification
NAT	*N*-acetyltransferase
NBS	*N*-bromosuccinimide
NBT	nitro-blue tetrazolium salt
NCL	native chemical ligation
NCS	*N*-chlorosuccinimide
NCS	noncrystallographic symmetry
NGS	next-generation sequencing
NHEJ	non-homologous end joining
NIR	near-infrared
NLS	nuclear localization signal
NMR	nuclear magnetic resonance
NOESY	*n*uclear *O*verhauser *e*ffect *s*pectroscop*y*
NP-LC	normal-phase liquid chromatography
NSCLC	non-small cell lung cancer
NSOM	near-field scanning optical microscopy
OD	optical density
OLA	oligonucleotide–ligation assay
OMA	optical multichannel analyzer
ONPG	*o*-nitrophenyl-β-D-galactopyranoside
OPA	*ortho*-phthaldialdehyde
OP-Cu	bis(1,10-*ortho*-phenanthroline)-copper(I) complex
ORD	optical rotation dispersion
ORF	open reading frame
PA	photoactivatable
PACE	programmable autonomously controlled electrode
PAcIFIC	precursor acquisition independent from ion count
PAD	pulsed amperometric detection
PAGE	polyacrylamide gel electrophoresis
PALM	photoactivated localization microscopy
PAM	protospacer adjacent motif
PAP	peroxidase–anti-peroxidase
PAT	process analytical technology
PBPC	polar bonded-phase chromatography
PBS	phosphate-buffered saline solution
PC	peak capacity
PCA	protein fragment complementation assay
PCA	principal component analysis
PCMD	power of combinatorial molecular discrimination
PCR	polymerase chain reaction
PDB	Protein Data Bank
PDMS	polydimethylsiloxane
PED	pulsed electrochemical detection
PEG	poly(ethylene glycol)
PELDOR	pulsed electron–electron double resonance
PFGE	pulsed-field gel electrophoresis
PID	photon-induced dissociation
PITC	phenyl isothiocyanate
PMF	peptide mass finger print
PMMA	poly(methyl methacrylate)
PMSF	phenylmethylsulfonyl fluoride
PNA	peptide nucleic acid
PNGase F	peptide *N*-glycosidase F
PPC	pressure perturbation calorimetry
PPI	protein–protein interaction
PQC	piezoelectric quartz crystals
PRE	paramagnetic relaxation enhancement
PRM	parallel reaction monitoring
PS	power spectrum
PSD	post source decay
PSF	point spread function
PSSM	position-specific scoring matrix
PTC	phenylthiocarbamoyl
PTH	phenylthiohydantoin
PTM	post-translational modification
PTR	proton transfer reaction
PVDF	poly(vinylidene fluoride)
PVP	polyvinylpyrrolidone
QCL	quantum cascade laser
QD	quantum dot
QSRR	quantitative structure–retention relationship
QY	quantum yield
R_S	resolution
RAM	restricted access material
RCF	relative centrifugal force
RCR	repair chain reaction
RDC	residual dipolar coupling
RF	replicative form
RFLP	restriction fragment length polymorphism
RGB	red-green-blue
RIA	radioimmunoassay
RISC	RNA-induced silencing complex
RMSD	root-mean-square deviation
RNAi	RNA interference
RNP	ribonucleoprotein
RP	reversed phase
RPA	ribonuclease protection assay
RPC	reversed-phase chromatography
RP-LC	reversed-phase liquid chromatography
rpm	rotations per minute
RRM	relative resolution map
rRNA	ribosomal RNA
RSV	Rous sarcoma virus
RSV	respiratory syncytial virus
RT	reverse transcriptase
S/MRM	selected or multiple reaction monitoring
SAA	serum amyloid A
SAXS	small-angle X-ray scattering
SBLD	structure-based ligand design
SCAM	scanning cysteine accessibility method
SCX	strong cation exchange
SD	standard deviation
SDA	strand displacement amplification
SDS	sodium dodecyl sulfate
SDS-PAGE	SDS-polyacrylamide gel electrophoresis
SEC	size-exclusion chromatography
SELEX	systematic evolution of ligands by exponential enrichment
SEM	scanning electron microscope
SEV	secondary electron multiplier

SFM	scanning force microscopy
SFX	serial femtosecond crystallography
SHAPE	selective 2′-hydroxyl acylation analyzed by primer extension
shRNA	short hairpin RNA
SI	secondary ion ionization
SICM	scanning ion conductance microscope
SIDT	single ion in droplet theory
SILAC	stable isotope labeling by amino acids in cell culture
SIM	single-ion monitoring
SIMS	secondary ion mass spectrometry
SIR	single isomorphous replacement
siRNA	small-interfering RNA
SMALDI	scanning microprobe MALDI
SNIM	scanning near-field infrared microscopy
SNOM	scanning near-field optical microscopy
SNP	single-nucleotide polymorphism
snRNA	small nuclear RNA
SNV	single-nucleotide variant
SOG	singlet oxygen generator
SPE	solid-phase extraction
SPFS	surface plasmon fluorescence spectroscopy
SPM	scanning probe microscope
SPR	surface plasmon resonance
SPRI	solid-phase reversible immobilization
SREBP	sterol response element-binding protein
SRM	single reaction monitoring
ss	single-stranded
SSB	single-stranded binding
SSCP	single-strand conformation polymorphism
SSIM	saturated structured illumination
STED	stimulated emission depletion
STEM	scanning transmission electron microscope
STET	singlet oxygen triplet energy transfer
STM	scanning tunneling microscope
STORM	stochastic optical reconstruction microscopy
STRP	short tandem repeat polymorphism
STS	sequence tagged site
SWATH-MS	sequential windowed acquisition of all theoretical fragment ion mass spectra
τ_c	rotational correlation time
TAD	topologically associating domain
TAE	tris acetate
TAP	tandem affinity purification
TBAB	tetrabutylammonium bromide
TBE	tris borate
TE	tris-HCl/EDTA
TEA	triethylamine
TEAA	triethylammonium acetate
TEM	transmission electron microscope
TEMED	*N*,*N*,*N*′,*N*′-tetramethylethylenediamine
TEV	tobacco etch virus
TFA	trifluoroacetic acid
TFO	triplex forming oligonucleotide
TGGE	temperature gradient gel electrophoresis
TIC	total ion current
TIR	total internal reflection
TIRFM	total internal reflection fluorescence microscopy
TIS	toponome imaging system
TLC	thin-layer chromatography
TLCK	L-1-chloro-3-(4-tosylamido)-7-amino-2-heptanone hydrochloride
T_m	melting point
TMA	transcription-mediated amplification
TMT	tandem mass tag
TOCSY	*to*tal *c*orrelation *s*pectroscop*y*
TOF	time-of-flight
TPA	1-oxyl-2,2,5,5-tetramethylpyrroline-3-acetylene
TPCK	L-1-chloro-3-(4-tosylamido)-4-phenyl-2-butanone
t_R	retention time
TRAP	*trp* RNA-binding attenuation protein
TRF	time-resolved fluorescence
TROSY	*t*ransverse *r*elaxation *o*ptimized *s*pectroscop*y*
TSH	thyroid-stimulating hormone
TSS	transcription start site
t_0	void time
UAS	upstream activating sequence
Ub	ubiquitin
UHPLC	ultrahigh pressure liquid chromatography
UNG	uracil-*N*-glycosylase
UV	ultraviolet
UV-DAD	UV-diode array detection
V_0	void volume
V_R	retention volume
VSV-G	vesicular stomatitis virus-Glycoprotein
W	heat
WLC	worm-like chain
YAC	yeast artificial chromosome
YFP	yellow fluorescent protein
ZMW	zero-mode waveguide
ZPCK	carbobenzoxy-L-phenylalanine-chloro-methyl ketone

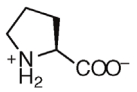
Appendix 3: Standard Amino Acids (three and one letter code)



C_α-substituted amino acids

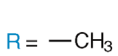


glycine (Gly) G



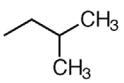
proline (Pro) P

non polar, aliphatic

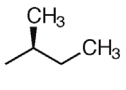


alanine (Ala) A

valine (Val) V

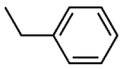


leucine (Leu) L

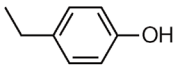


isoleucine (Ile) I

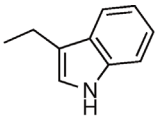
aromatic



phenylalanine (Phe) F



tyrosine (Tyr) Y

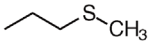


tryptophan (Trp) W

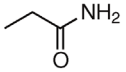
polar, uncharged

serine (Ser) S threonine (Thr) T

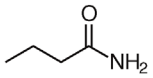
cysteine (Cys) C



methionine (Met) M



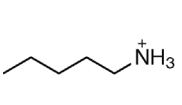
asparagine (Asn) N



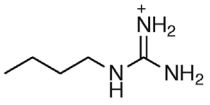
glutamine (Gln) Q

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

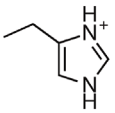
positively charged



lysine (Lys) K

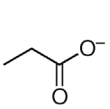


arginine (Arg) R

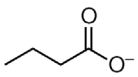


histidine (His) H

negatively charged



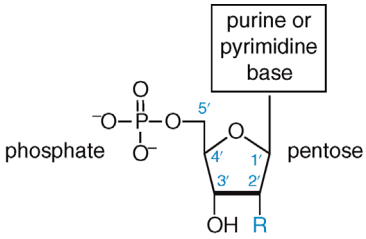
aspartic acid (Asp) D



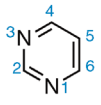
glutamic acid (Glu) E

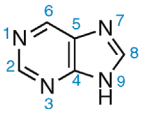
Appendix 4: Nucleic Acid Bases

General structure of nucleotides:



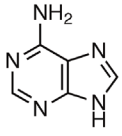
R = OH: ribose
H: deoxyribose

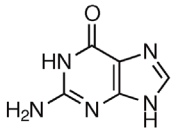

pyrimidine

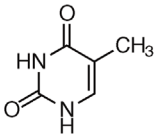

purine

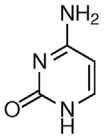
nucleoside: base + (deoxy)ribose
nucleotide: base + (deoxy)ribose + phosphate

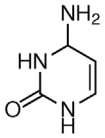
Nucleic acid bases


adenine
A


guanine
G


thymine
T
(DNA)


cytosine
C


uracil
U
(RNA)

purines **pyrimidines**

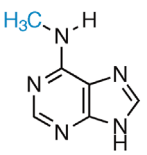
nucleosides:

DNA: deoxyadenosine	deoxyguanosine	(deoxy)thymidine	deoxycytidine	deoxyuridine
RNA: adenosine	guanosine	thymidine		uridine

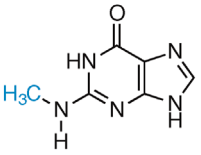
Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

Modified nucleic acid bases

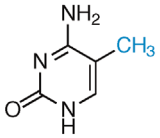
rare bases in DNA



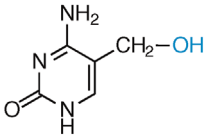
N^6-methyladenine



N^2-methylguanine

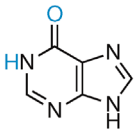


5-methylcytosine

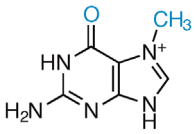


5-hydroxymethylcytosine

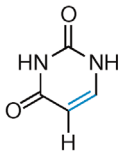
rare bases in RNA



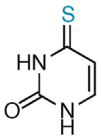
hypoxanthine –
as nucleoside: inosine



7-methylguanine



pseudouracil



4-thiouracil

Index

Bioanalytics: Analytical Methods and Concepts in Biochemistry and Molecular Biology, First Edition.
Edited by Friedrich Lottspeich and Joachim Engels.

G

H

I

J

K

L

M

O

Q

S

Y

Z

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.