Marco F. Schmidt

Chemical Biology

and Drug Discovery

Springer

Chemical Biology

Marco F. Schmidt

Chemical Biology

and Drug Discovery

Springer

Marco F. Schmidt
Potsdam, Germany

ISBN 978-3-662-64411-9 ISBN 978-3-662-64412-6 (eBook)
https://doi.org/10.1007/978-3-662-64412-6

This book is a translation of the original German edition „Chemische Biologie“ by Schmidt, Marco F., published by Springer-Verlag GmbH, DE in 2020. The translation was done with the help of artificial intelligence (machine translation by the service DeepL.com). A subsequent human revision was done primarily in terms of content, so that the book will read stylistically differently from a conventional translation. Springer Nature works continuously to further the development of tools for the production of books and on the related technologies to support the authors.

This Springer imprint is published by the registered company Springer-Verlag GmbH, DE part of Springer Nature.
The registered company address is: Heidelberger Platz 3, 14197 Berlin, Germany

Preface

This book has not been designed as a classical textbook. The focus is not the pure knowledge transfer of the results of all chemical-biological work (which would also go beyond the scope of a book). The focus is on application. Through the specific selection of chemical-biological techniques and concepts, an attempt is made to provide the reader with the necessary tools to be able to develop new ways of thinking in the complex field of chemical biology in drug development and thus, hopefully, new therapeutic options. The stated aim of this book is to provide students, (post-)graduate students, and experienced scientists at universities as well as in industry with concrete solutions to their problems and to inspire them.

After an introduction to the problems that chemical biology addresses in drug discovery, the three levels of molecular biology dogma: DNA, RNA, and proteins and their role as drug targets serve as the common thread of this book:

Part I: Introduction concretizes the problem of developing drugs for genes newly identified in genomics. It introduces the subfields of chemical biology: Drug and drug target discovery of chemical genomics, drug target characterization of chemical proteomics, and drug target validation of chemical genetics, and how these are combined to form chemical biology.

Part II: DNA focuses on genome research. Starting with the role of genome-wide association studies to identify disease-relevant genes, DNA as a drug target and ending with gene editing and gene therapies. This part takes up the methods currently associated with very high hopes, such as artificial intelligence in the analysis of genome data and the gene-editing method CRISPR/Cas as the key to new therapies.

Part III: RNA deals with the role of RNAs as new drug targets. Since many proteins cannot be manipulated in their function with a drug, drug developers are increasingly focusing on RNA molecules. RNA drug targets in translation or gene regulation (e.g., RNA interference: RNAi) offer unexpected possibilities. Arranged according to RNA subgroups, the challenges and potentials are discussed using successful case studies.

Part IV: Peptides and *proteins* represent the central part of this book. Proteins are the functional carriers in living systems. Modulation of their functions by a drug is the central concept of drug discovery and drug development. Most drugs bind to proteins. After a brief introduction to the structure, function, and synthesis of peptides and proteins, chemical genomics, chemical proteomics, chemical genetics, and chemical biology of proteins are described in detail.

Who Is the Book for?

The book incorporates the new challenges in drug discovery that chemical biology is trying to answer. For the first time, the reader is provided with a navigator through chemical biology that focuses on the new approaches to drug discovery from gene to the drug. In short, it covers the *must-know* for current and future pharmaceutical researchers and is aimed at both undergraduate and (post-)graduate students in the life sciences as well as experienced scientists in the pharmaceutical industry.

Potsdam, Germany Marco F. Schmidt

Contents

Part III RNA

Part IV Proteins

Introduction 1

Chemical biology is not a new science, although the term was only formulated in the 1990s. The first chemical biology experiments in which chemical probes were used to study biological systems date back almost 200 years. Nevertheless, it was not until the end of the twentieth century that chemical biology was given its name in this sense, and since then it has begun an unprecedented triumphal march worldwide at universities as well as in the pharmaceutical industry. Today, courses on chemical biology are offered at almost all universities, and furthermore, in many universities chairs have been established or institutes and faculties have been created. For example, the Department of Chemistry at Harvard University has expanded to become the Department of Chemistry & Chemical Biology. Meanwhile, all major pharmaceutical companies have now established chemical biology departments in addition to their established drug discovery research departments. This does not even take into account the countless biotech start-ups in the field of chemical biology in the last 20 years. What has led to the fact that chemical biology has become indispensable at universities and, above all, in the industry in such a short time?

The reasons for this development lie on the one hand in the technical possibilities and on the other hand in economic need. It was not until the end of the twentieth century that both in vitro and in vivo experiments of chemical probes could be better observed, or observed at all, with the help of new bioanalytical techniques, such as spectroscopy or microscopy. Before this, it was not practical to systematically bioanalyze drugs because the changes triggered could not be detected in living organisms, such as a cell. In addition, there is an economic reason for the rise of chemical biology. The approval requirements for new drugs have been successively heightened by regulatory authorities since the 1960s. In the past, drugs were approved without knowing their mechanism of action. In some cases, this resulted in serious side effects and even deaths after approval. A well-known example in Germany is the thalidomide scandal. This forced the pharmaceutical industry to develop (or adopt) new techniques that could clarify the interactions of their drug candidates with

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_1

biomolecules—and what clinical consequences these interactions could have—before clinical testing in the organism. In summary, in the 1990s, new bioanalytical technologies met the pharmaceutical industry's need to more carefully pre-test their drug candidates for safety—with the hope of increasing success rates in expensive clinical trials. Nevertheless, these reasons do not fully explain why chemical biology plays a decisive role in drug development today.

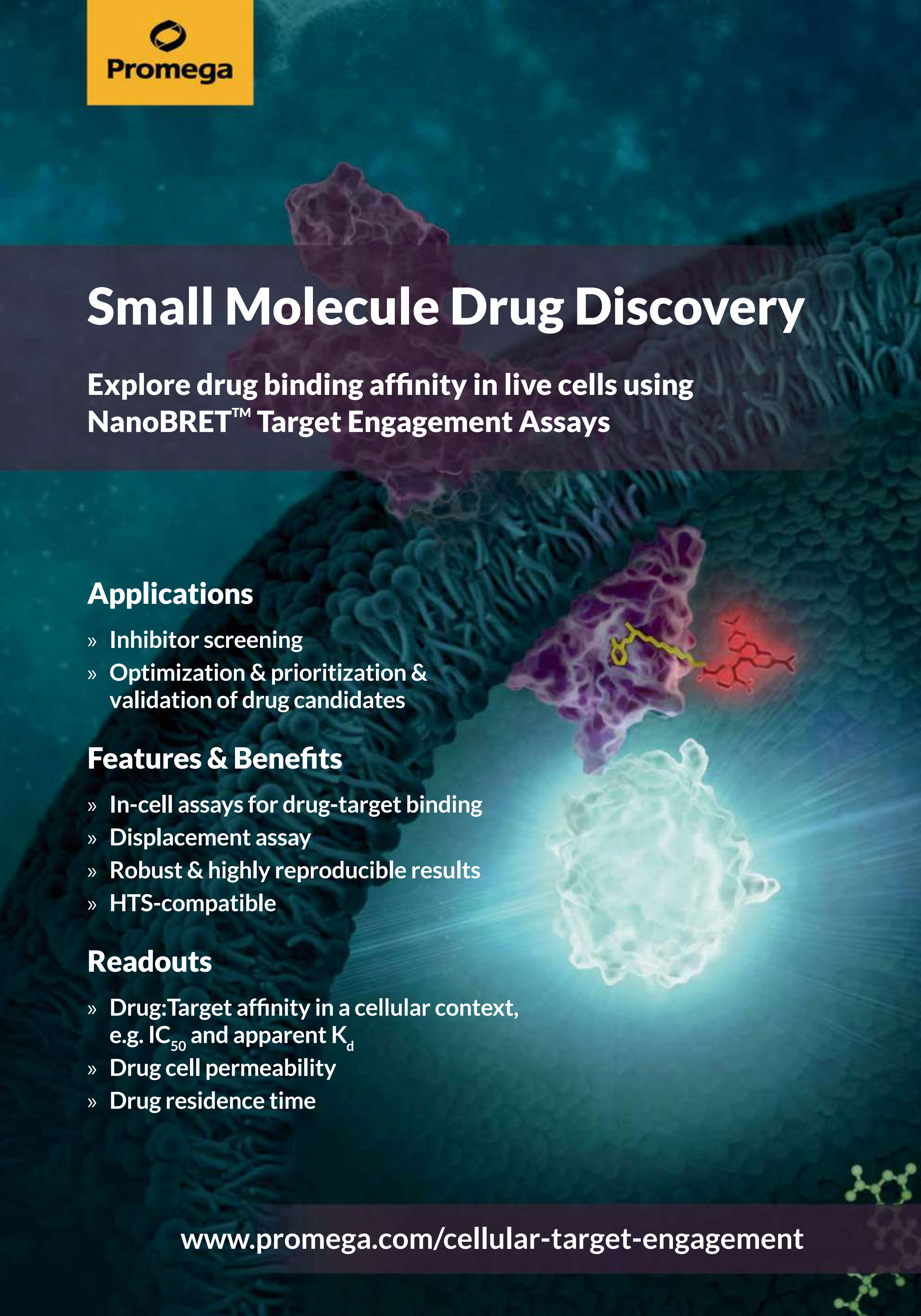
In 2001, when parts of the human genome were first published, the scientific director of the Human Genome Project, Francis Collins, declared, "The human genome is a revolutionary textbook of medicine with insights that give health care providers immense new powers to treat, prevent and cure disease." The statement reflects the hopes of the time that advances in molecular biology would usher in a new era of medicine. In contrast, 20 years later, it is unfortunate to note that the sequencing of the human genome has not had the hoped-for effect of enabling the development of a large number of new drugs in a very short time. Why is this?

While genomic research has indeed found a large number of disease-relevant genes, the identification of a disease-relevant gene does not make a drug. It quickly became apparent that most of the proteins encoded by the identified genes do not bind any chemical probe due to their structure and thus cannot be modulated in their function. This is referred to as the *ligandability of* a macromolecule. Furthermore, the binding of an chemical probe to a gene product does not inevitably lead to a clinical effect. Proteins of disease-relevant genes associate with the disease but do not have to occupy the central position in the signaling pathway, and so intervention with an agent does not necessarily result in a clinical effect. The ability to use a drug to manipulate the function of a gene product to produce a clinical effect is known as *drugability*. The challenge of investigating the ligand binding and drugability of gene products is taken up by chemical biology in drug development.

In summary, chemical biology in drug development closes the gap between the results of modern genomic research, such as the identification of disease-relevant genes on the one hand, and the social and economic demand to develop new drugs from this knowledge on the other. This ultimately explains the enormous importance of chemical biology for drug development in the age of genomic research.

Part I

Chemical Biology and Drug Discovery



Promega

Small Molecule Drug Discovery

Explore drug binding affinity in live cells using NanoBRET™ Target Engagement Assays

Applications

» Inhibitor screening

» Optimization & prioritization & validation of drug candidates

Features & Benefits

» In-cell assays for drug-target binding

» Displacement assay

» Robust & highly reproducible results

» HTS-compatible

Readouts

» Drug:Target affinity in a cellular context, e.g. IC_{50} and apparent K_d

» Drug cell permeability

» Drug residence time

www.promega.com/cellular-target-engagement

Chemical Biology: A Holistic Science 2

Chemical biology is the science that uses ***chemical*** substances, usually synthetically produced, to understand and manipulate complex ***biological*** systems.

Chemical biology is a scientific discipline that encompasses the fields of chemistry and biology. The discipline involves the application of chemical techniques and analysis—as well as the use of chemical compounds, usually organically synthesized—to study and manipulate biological systems. In contrast to biochemistry, the chemistry of biology, which involves the study of biomolecules and the regulation of biochemical pathways within and between cells, chemical biology deals with the **application of** chemistry ***to*** biology.

Chemical biology is an interdisciplinary science. Furthermore, it distinguishes itself from reductionist sciences, which attempt to describe composite systems as the sum of their components and their properties. In contrast, chemical biology as a new scientific discipline follows the approach of scientific holism i.e. the assumption that biological systems and their properties should be considered as a whole and not as a simple addition of their parts. Chemical biology is scientifically, historically, and philosophically rooted in bioorganic chemistry, medicinal chemistry, and pharmacology as well as supramolecular chemistry and especially genetics, molecular biology, and biochemistry. Chemical biology, therefore, draws on many methods of the aforementioned sciences.

2.1 Reductionism and Holism in the Life Sciences

Nevertheless, against this background, chemical biology differs considerably in its holistic approach from its roots in the classical, reductionist life sciences, such as genetics, molecular biology, and biochemistry, where it is assumed that biology can be reduced to chemistry and chemistry to physics. For example, in biochemical, genetic, or molecular biology experiments, new observable changes (phenotype) in gene products are introduced

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_2

into organisms, cells, or biomolecules by creating mutations in genes (mutagenesis) and are studied intensively. This presupposes, in a reductionist sense, that the system as a whole can be influenced by the alteration of a single element: Mutagenesis of a gene alters the blueprint of the organism (see Fig. 2.1). However, the totality of the organism is ignored. Phenotypic change may only be the consequence of gene mutagenesis depending on certain environmental factors, or the network may compensate for interruptions in a signaling pathway due to mutagenesis. This is where the philosophical dividing line between the new discipline of chemical biologyand the other biological sciences emerges. Living systems, such as organisms or cells, are studied as a whole. Instead of mutating individual genes, attempts are made to *reversibly* manipulate the function of genes and gene products by synthetically produced compounds in in vitro or in vivo experiments and to investigate the changes in the system as a whole. Only by accepting a living system with its properties as a whole and not as the sum of its parts is it possible to study more than the sum of all parts of life (see Table 2.1).

In order to be able to reversibly manipulate the function of genes and gene products, synthetically produced compounds, so-called *probes*, are used. Figure 2.2 shows the three classes of macromolecules from the central dogma of molecular biology—DNA, RNA, and proteins—along with which synthetic molecules or procedures can modulate their functions. The central dogma of molecular biology describes the information transfer within biopolymers from DNA to RNA to proteins. Proteins, as classical drug targets, can be influenced in their function by small molecules, peptides, or antibodies. RNA molecules also bind directly to small molecules (RNA binders) and all variants of short nucleotide sequences (oligonucleotides). In addition, the function of RNA can be modulated by influencing RNA interference (RNAi). DNA can be influenced in its function as an information carrier via DNA binders, or more broadly via knock-out or genome editing (genome rewriting) with CRISPR-Cas, for example.

Manipulating genes and gene products with chemical probes is one thing. The other is to be able to observe phenotypic change. Chemical biology is, therefore, inextricably linked to the improved bioanalytical methods of recent years.

However, it is not only the above-mentioned current bioanalytical methods that are different. Chemical biology is highly automated. Chemical probes are not individually tested by hand. There are now large libraries of thousands to millions of potential chemical probes all over the world that are tested in *high-throughput* robots in miniaturized biochemical experiments, so-called *assays*, in the shortest possible time.

Despite the bioanalytical methods that are only possible today, the idea of using organic compounds to examine cells, tissue samples, plants, insects, or even animals is almost as old as organic chemistry. In 1856, the British chemist William Perkin (1838–1907), who was only 18 years old, accidentally discovered the first aniline dye while trying to synthesize the antimalarial agent quinine. He named this dye Mauveine because of its violet color, similar to the flower of the wild mallow, and in reference to the French name Mauve (see Fig. 2.3). The discovery of other aniline dyes followed. Initially, aniline dyes produced industrially in large quantities met the needs of the textile industry. The

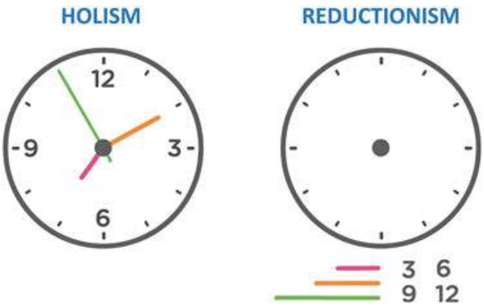


Fig. 2.1 Comparison of the approaches of reductionism (component-based) and holism (system-based)

Table 2.1 Reductionism and holism in the life sciences

Reductionism	Holism
A complex system can be understood by studying its components	The principle of a higher order cannot be explained meaningfully by testing the individual components in isolation
Example: The role of DNA in heredity has been deduced by studying its molecular structure	Example: A cell disassembled according to its chemical components is no longer a cell. It is also difficult to analyze a complex process without disassembling it

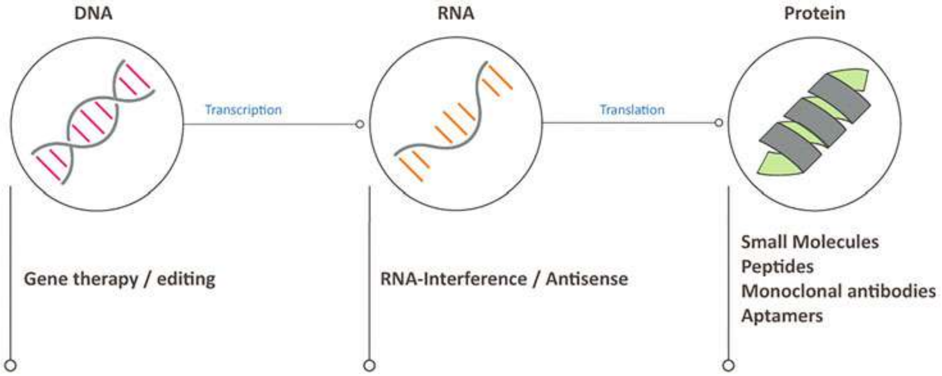


Fig. 2.2 The central dogma of molecular biology describes the transfer of information from DNA via RNA to the functional carrier protein, which is determined by the order (sequence) of the respective monomers (nucleotides in the case of DNA and RNA; amino acids in the case of proteins). The function of the three biopolymers can be influenced by chemical compounds or by biochemical processes in chemical-biological experiments

production of naturally occurring plant dyes, such as indigo for dyeing cotton fabrics was tedious, laborious, and ultimately very expensive. Colored clothing was a luxury item. The discovery of the inexpensive production of synthetic aniline dyes from coal tar led to the emergence of the modern chemical industry. The best example is the world’s largest

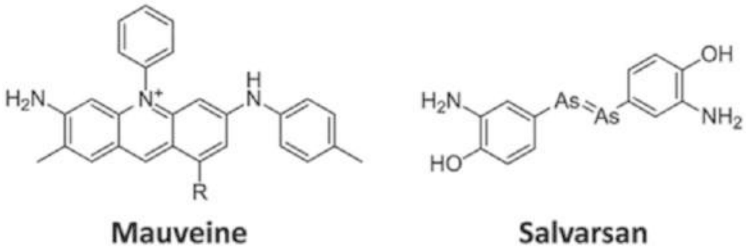


Fig. 2.3 Structures of the first aniline dye mauveine (R = mixture of CH_3 and H) produced synthetically by William Perkin and of the first chemotherapeutic agent salvarsan discovered by Paul Ehrlich for the treatment of the infectious disease syphilis

chemical company at present, Badische Anilin- & Soda-Fabrik, better known as BASF. Similarly, the Actien-Gesellschaftfür Anilin-Fabrication, i.e. Agfa, is known today for its photographic products.

The idea soon arose that aniline dyes, which were found in the attempt to produce an antimalarial agent, could themselves have pharmacological properties. The best-known example is the work of Paul Ehrlich (1854–1915). Ehrlich used aniline dyes to stain cells in his early research. This enabled the diagnosis of numerous blood diseases for the first time. He suspected that the different staining of cells was due to chemical reactions with individual components of the cells. He then postulated his hypothesis of *magic bullets*. Inspired by the opera “Der Freischütz” (“The Freeshooter”), he chose the name “magic bullets” for chemical substances which, like in the opera, always hit the pathogen with pinpoint accuracy and spare healthy tissue *(“chemotherapiaspecifica”)*. In 1909, Ehrlich succeeded in discovering a chemical that selectively killed syphilis pathogens. The compound was named salvarsan, in reference to the Latin words *salvare*, meaning to heal, and *sanus*, meaning healthy, as well as a fragment of the word arsenic: “To heal with arsenic” (see Fig. 2.3). It is often referred to as *preparation 606*, as it was found in the 606th animal experiment. Salvarsan turned out to be a milestone in drug research. For the first time in medicine, it was possible to use a targeted antimicrobial drug against an infectious disease without severe side effects, such as hair loss or tooth loss, as was the case with the mercury therapy used until then. Ehrlich is, thus, considered the founder of modern chemotherapy.

2.2 Chemical Biology Is Not Drug Development

Nevertheless, it is important to emphasize that chemical biology is not synonymous with, nor is it a sub-discipline of, drug discovery and drug development! The aim of chemical biology is not to cure or treat a disease but to study living systems with chemical probes. These probes do not meet the high health and safety standards to which chemical probe and especially drugs are subject. Many toxicity issues can be ignored as these compounds are not intended for human use. On the contrary, probes with reactive chemical functions, such as Michael acceptor, epoxides, chloromethylene ketones, etc.—which are avoided in drug discovery—are used in chemical biology to achieve stronger biological effects in experiments. However, the use of probes with reactive groups carries the risk of side effects, so-called off-target effects, which are the consequence of the specific binding of the probes to other drug targets.

Despite this fundamental difference, chemical biology is very closely linked to drug development. Methodologically, chemical biology is almost indistinguishable from drug development: Chemical compound libraries are screened for phenotypic changes in high-throughput procedures and subsequently applied in biochemical or cell biological experiments. However, the chemical probes used in chemical biology substance libraries are usually not compounds suitable for drug development. Nevertheless, chemical probes serve as concept compounds for drug discovery. The application of chemical probes serves to elucidate the complexity of living systems with the consequence of developing new drug targets and concepts of action.

Summary

Chemical biology has its roots in the biological sciences. However, it differs significantly in its approach to the holistic study of living systems. It is a holistic science. In contrast, the classical disciplines of the natural sciences follow a reductionist approach: Biology can be reduced to chemistry and chemistry to physics.

Chemical biology is not the same as (nor a part of) drug discovery or pharmacology. Chemical biology only uses the same methods of screening chemical compound libraries to find chemical probes for the study and manipulation of living systems. Chapter 3 explains why chemical biology is nevertheless now essential for drug discovery.

Further Reading

Crick F (1970) Central dogma of molecular biology. Nature 227:561–563

Mayr E (1982) The growth of biological thought: diversity, evolution, and inheritance. Harvard University Press, Boston

Morrison KL, Weiss GA (2006) The origins of chemical biology. Nat Chem Biol 2:3–6

Schreiber SL (2005) Small molecules: the missing link in the central dogma. Nat Chem Biol 1:64–66

Van Regenmortel MHV (2004) Reductionism and complexity in molecular biology. EMBO Rep 5: 1016–1020

3 Drug Development

The development of a new drug is lengthy, risky, and therefore expensive. According to studies by the management consultancy firms PricewaterhouseCoopers (PwC) and Deloitte, the expenditure of the 12 largest pharmaceutical companies on drug development has been rising for years. In its report “From Vision to Decision Pharma 2020”, PwC reports that the average cost of a new drug now exceeds four billion USD. At the same time, the number of new drug approvals by the US regulatory agency, the *Food and Drug Administration* (FDA), is stagnating, as Fig. 3.1 shows. Since 2010, the number of new drug approvals has leveled off at 30 per year. Based on the trend of stagnant approval numbers with rising costs, Deloitte calculated that the return on investment, or profit relative to the capital employed, of development investments in the pharmaceutical industry, has declined from 10.1% in 2010 to 3.2% in 2017. A return on investment of permanently less than 3%, after deducting inflation (the depreciation of the value of a currency or the rate of inflation) of 2% on average since 2000, would mean that there is no money to be made from the development of new drugs. The social consequence of this economic development would be catastrophic: No more new therapies would be developed.

Nevertheless, the figures mentioned should be viewed critically: Patient associations, as well as health insurance companies, doubt the figures mentioned. On the basis of publicly available annual financial statements of listed pharmaceutical companies, these organizations have calculated the actual costs for the new development of a drug at 300 to 700 million euros. This is significantly lower than the industry’s figures. Pharmaceutical companies justify the high prices for new drugs with the rising development costs. Accordingly, the pharmaceutical industry has an interest in citing high costs in reimbursement discussions with health insurers.

Despite all the relativization of these figures, it must be noted that the development of new drugs is stagnating. If we compare drug development with information technology,

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_3

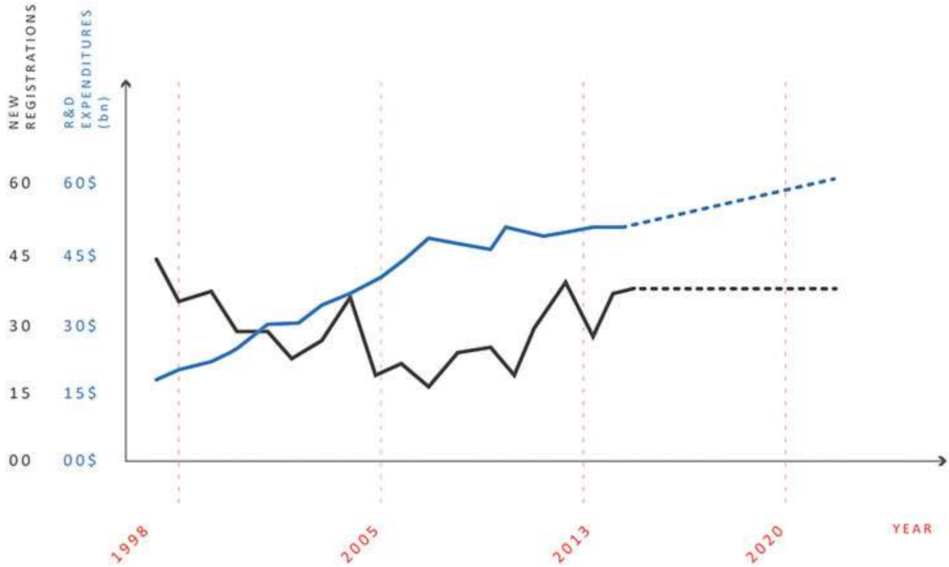


Fig. 3.1 Evolution of expenditure (in blue) on drug development by the 12 largest pharmaceutical companies and the number of new drug approvals (in gray) by the US regulatory authority FDA. The forecast evolution up to the year 2025 (dashed in each case) assumes a further increase in expenditure with stagnation of new approvals

this becomes particularly clear: Gordon Moore postulated in 1965 that the number of transistors on a microprocessor, and thus its computing power, doubles every 2 years (Moore's Law). Although enormous technological leaps have been made in the last 50 to 100 years, we cannot develop drugs better or faster. With the sequencing of a human genome for the first time in 2001, the idea arose that similar technological leaps could be expected in drug development. Against this backdrop, the head of the Human Genome Project at the time, Francis Collins, claimed that the human genome gives healthcare providers new powers to treat, prevent and cure disease. The cost of sequencing a human genome has dropped from nearly 100 million USD in 2001 to 1000 USD in 2017. This development beats Moore's Law by far, which "only" assumes a doubling of effectiveness every 2 years (see Fig. 3.2). Although the enormous cost reduction has enabled millions of human genomes to be sequenced, the hoped-for effect on drug development has failed to materialize. What is the reason for this?

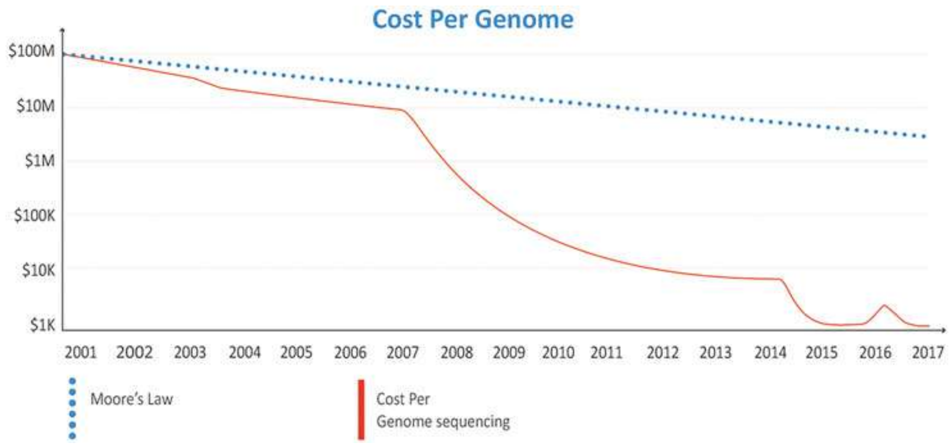


Fig. 3.2 Cost of genome sequencing since 2001 compared to Moore's law. (Source: www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data)

3.1 Drug Development

The development of a new drug can be divided into several stages (Fig. 3.3): The initial search for chemical probes with subsequent lead structure development, the preclinical studies, the clinical studies with their three phases, and the review process by the authorities (usually the US FDA).

3.1.1 Search for Hit Molecule

In the first stage of drug development (Fig. 3.4), a hypothesis is developed with the aid of known data. This means that a macromolecule, a so-called drug target, is investigated by manipulating its function (e.g., by genetic manipulation = mutagenesis) to see whether modulation of the drug target causes a phenotypic change in cell or animal experiments. After a large number of these experiments with different drug targets, a decision is made and a hypothesis is defined that a disease can be treated, perhaps even cured, by modulating a particular drug target. This is called the drugability of a drug target. The search then begins for a suitable compound that binds to the drug target and can reproduce the effect in cell and animal experiments. For this purpose, large substance libraries with millions of compounds are screened against the drug target. This process is known as high-throughput screening. In most cases, this is the first hurdle because many promising drug targets do not bind or bind poorly to potential drugs. Ligandability is the ability of a macromolecule to bind a potential chemical probe (ligand).

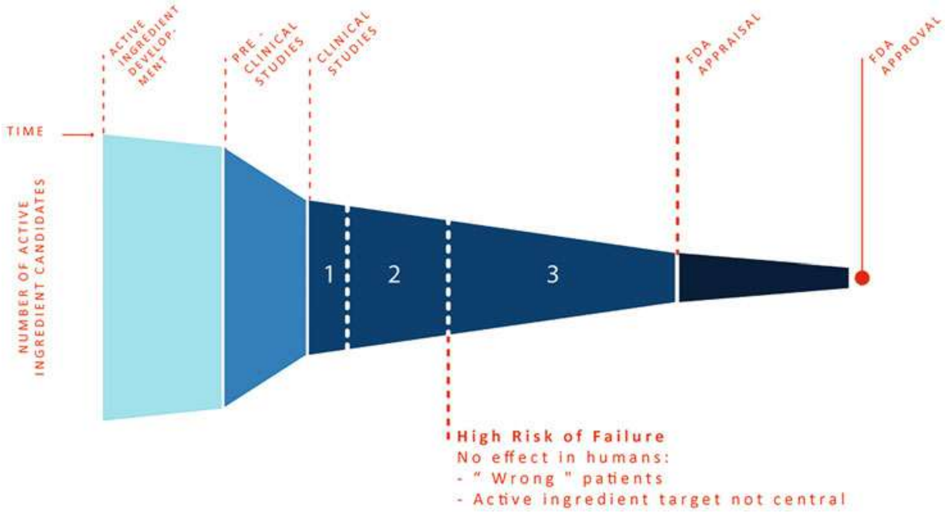


Fig. 3.3 The development process of a new drug

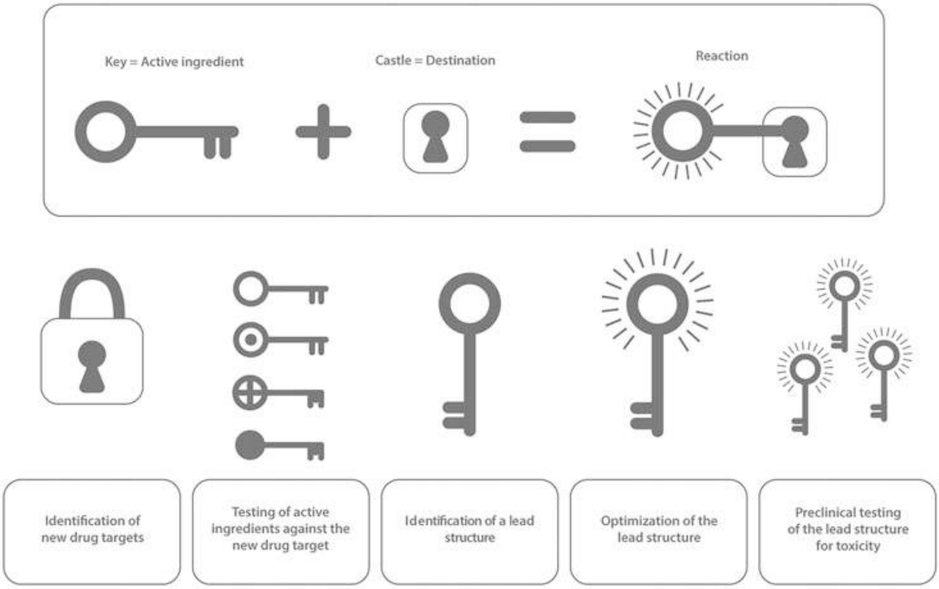

Key = Active ingredient
Castle = Destination
Reaction
Identification of new drug targets
Testing of active ingredients against the new drug target
Identification of a lead structure
Optimization of the lead structure
Preclinical testing of the lead structure for toxicity

Fig. 3.4 Schematic representation of drug discovery, lead development, and preclinical testing based on the lock-and-key principle. Once a potential drug target = lock (usually a protein) has been identified, a drug = key is sought that binds to it. Often, the chemical probe that is identified must be further optimized in order to finally be tested (pre)clinically

3.1.2 Lead Structure Development and Preclinical Studies

Once a ligand has been found, a large number of derivatives (structurally similar compounds) are produced in the hope of optimizing binding properties, such as affinity and selectivity. This process is called lead structure development. It is a central part of medicinal chemistry (formerly known as pharmaceutical chemistry). In addition to affinity and selectivity, attention is already paid to avoiding the toxic properties of the ligand. Once a promising lead structure has been identified, preclinical studies are carried out. The focus here is on toxicological testing in animal experiments.

In most cases, lead development and preclinical studies run parallel to each other in order to identify any toxic effects of a lead early on.

3.1.3 Phase I Clinical Trial

If the chemical probe has been tested toxicologically and found harmless in animal experiments, clinical testing is carried out in humans. In clinical phase I, the toxicological tests are repeated on humans instead of animals. For this purpose, different doses of the chemical probe are tested in mostly healthy volunteers. Here, particular attention is paid to ensuring that the results from the animal tests can be reproduced in humans. If the chemical probe is toxicologically safe for humans in the context of its application, clinical phase II follows.

3.1.4 Phase II Clinical Trial

This is the first time that therapeutic efficacy in humans is considered. Therefore, the studies are carried out on patients. The aim of the phase II clinical trial is always to find the optimal dosage.

3.1.5 Phase III Clinical Trial

Once the optimal dosage has been found in phase II, a phase III clinical trial follows. Here, the decisive proof of efficacy must be provided. Phase III studies are usually randomized double-blind studies. This means that the patient is randomly assigned to receive the chemical probe or a placebo. Neither the patient nor the treating physician knows which drug they will receive (hence, *double*-blind).

3.2 Reasons for the Failure of Drug Development

Studies of failed drug development efforts from 2012–2015 have shown that success rates vary by stage of development. The highest failure rates occur in clinical phase II. The success rate in this phase is only 25%.

	Preclinic	Phase I	Phase II	Phase III	Expert opinion	Total
Success rate	65%	45%	25%	65%	85%	10%

What are the reasons for failure and why is clinical phase II the time with the highest failure rate?

The reasons for the failure of drug development efforts are shown in Fig. 3.5.

Accordingly, most drug developments fail due to a lack of efficacy. This explains why clinical phase II has the lowest success rate. In this phase, the efficacy of the candidate drug is tested for the first time. Why is it not possible to predict clinical efficacy based on laboratory experiments so as not to experience a surprise in clinical testing?

3.3 The Gap Between Laboratory Experiments and Clinical Effect

Drug discovery is a hypothesis-testing process. An idea of how a disease can be treated is developed using cell and animal experiments at the beginning of a hypothesis. This is tested in humans through a variety of experiments up to final trials. Does anything strike the reader here? The process is reductionistic. What works in individual experiments in cells and animals should work on a large scale in humans. However, drug effects in humans are more than the sum of results from cell and animal experiments. The data from these experiments is too weak to reliably predict the clinical efficacy of a drug. This is evident from studies on the failure of phase II and phase III clinical trials. In particular, phase II clinical trials are the first to test for efficacy. This circumstance is also referred to as the gap between laboratory experiments and clinical efficacy.

The reasons are that the selected drug target is not central to the disease mechanism and that patients were selected for the study who have the clinical symptoms of the addressed disease but do not suffer from this disease. The latter sounds trivial, but it is a major problem. Several trials of the efficacy of a drug for treating the late form of Alzheimer's disease have been negative. After autopsies of deceased participants were conducted, it was determined that up to 35% of the participants were clinically demented but did not have Alzheimer's disease. That we cannot predict whether—and especially in whom—compounds will be clinically efficacious, despite genomics research, is one reason why drug development is risky. But there is another reason, which is rooted in the concept of how we develop drugs.

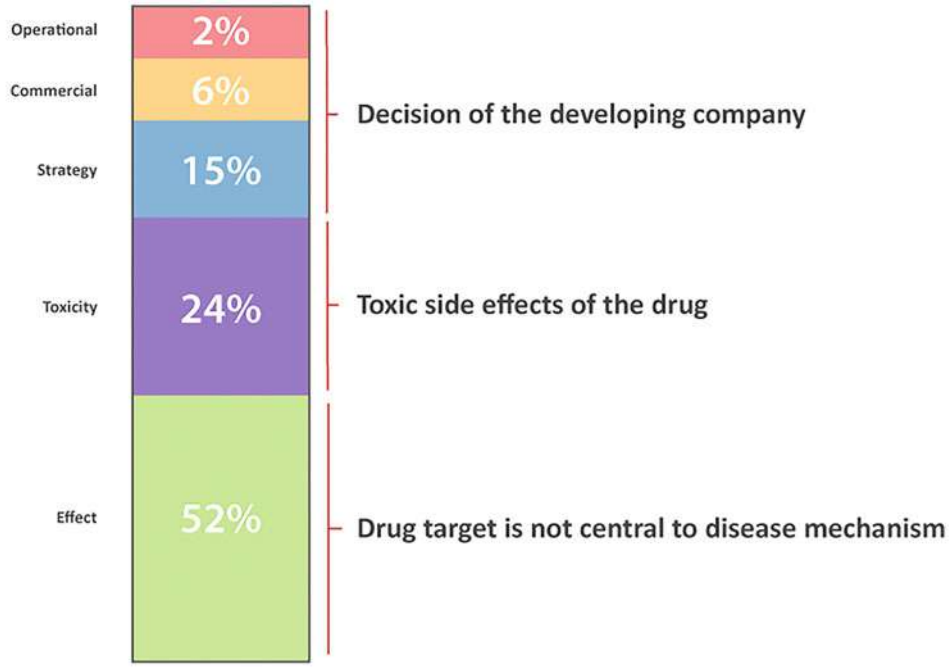


Fig. 3.5 Reasons for the failure of drug development efforts

Summary

Drug development is a holistic discipline, contrary to the methods used in the reductionist biosciences. This is problematic in that one cannot predict clinical efficacy in humans based on laboratory and cellular experiments. The latter leads to appallingly low success rates when the drug is first tested for efficacy in humans in phase II clinical trials. The pharmaceutical industry's return on investment has been in decline for years. This means that in the future it will no longer be possible to profit from the development of new drugs. The social consequence of this economic development would be catastrophic: No new therapies would be developed.

Further Reading

Harrison RK (2016) Phase II and phase III failures: 2013–2015. Nat Rev Drug Discov 15:817–818

https://www.pwc.com/gx/en/pharma-life-sciences/pharma2020/assets/pwc-pharma-success-strategies.pdf

https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Life-Sciences-Health-Care/deloitte-uk-measuring-return-on-pharma-innovation-report-2018.pdf

https://www.mckinsey.com/~/media/mckinsey/industries/pharmaceuticals%20and%20medical%20products/our%20insights/precision%20medicine%20opening%20the%20aperture/precision-medicine-opening-the-aperture.ashx

4 From Genomic Research to Chemical Biology

Another reason why the development of new drugs is difficult is grounded in genomic research. The genome codes for about 23,000 genes, that is, sites that lead (via *transcription*) to the production of biologically active ribonucleic acids (RNA), which in turn code for 50,000 to 100,000 proteins (via *translation*). In light of these numbers, currently approved drugs were examined. As Table 4.1 shows, 1591 drugs are currently approved by the US regulatory authority (FDA). Where possible, each drug was assigned to a drug target. Of the 1591 approved drugs, 1292 bind to 695 human drug targets. The drugs, therefore, target only 2–3% of the human genome (Fig. 4.1). Looking at the numbers more closely, it is evident that 1348 of the 1591 approved substances are low molecular weight ligands—chemical compounds with a molecular mass of less than 1000 g/mol—which bind to 749 of the 893 identified drug targets. An average drug target is a low molecular weight ligand that binds to a protein. The protein drug targets are almost exclusively enzymes, channels, receptors, and transporters. They have in common the property of binding a low molecular weight ligand (i.e., *ligandability*). In most cases, this is because the afore mentioned protein classes bind a natural ligand, for example, a substrate. The proteins have defined binding pockets in which natural ligands, as well as chemical probes, can bind very well. Enzymes, channels, receptors, and transporters are signal amplification points. Chemically, these proteins do not occur in signaling pathways in stoichiometric amounts, but in catalytic concentrations. They catalyze biochemical signaling steps. The elimination of such a step prevents signal amplification. A clinical effect (such as the effect of a drug) occurs. A drug target's propensity to undergo modulation that triggers a clinical effect is called its *drugability*. For these two reasons, drug development has historically focused on small molecule ligands that bind to proteins such as enzymes, channels, receptors, or transporters. What do you do when you find a disease gene that doesn't code for one of these proteins?

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_4

Table 4.1 List of chemical probes approved as drugs by the US regulatory authority (FDA) and their targets

	Drug targets			Approved chemical probes		
	Total	Low molecular weight ligands	Biologics	Total	Low molecular weight ligands	Biologics
Human proteins	667	549	146	1194	999	195
Pathogenic proteins	189	184	7	220	215	5
Other human macromolecules	28	9	22	98	63	35
Other pathogenic macromolecules	9	7	4	79	71	8
Total	893	749	179	1591	1348	243

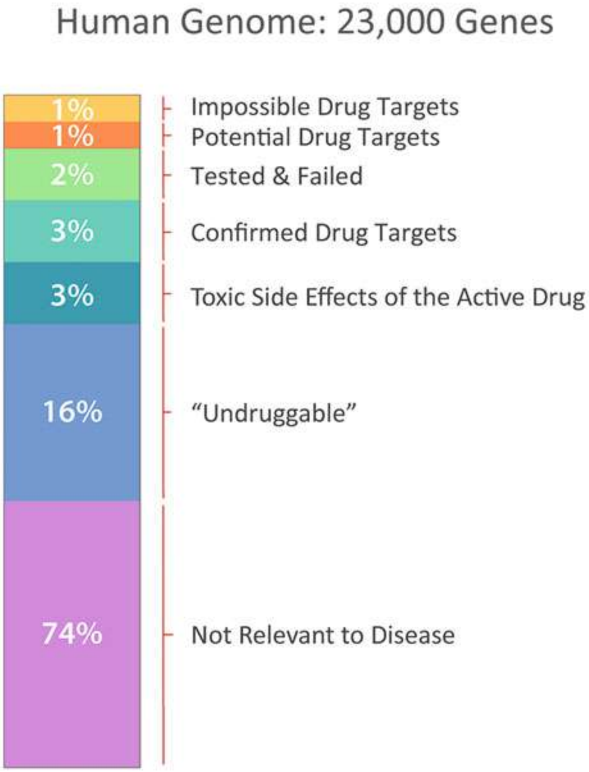


Fig. 4.1 The genome codes for approximately 23,000 genes. Of the 1591 approved drugs, 1292 bind to 695 human drug targets. Accordingly, the drugs target only 2–3% of the human genome. Most gene products are either not disease-relevant or are "undruggable"

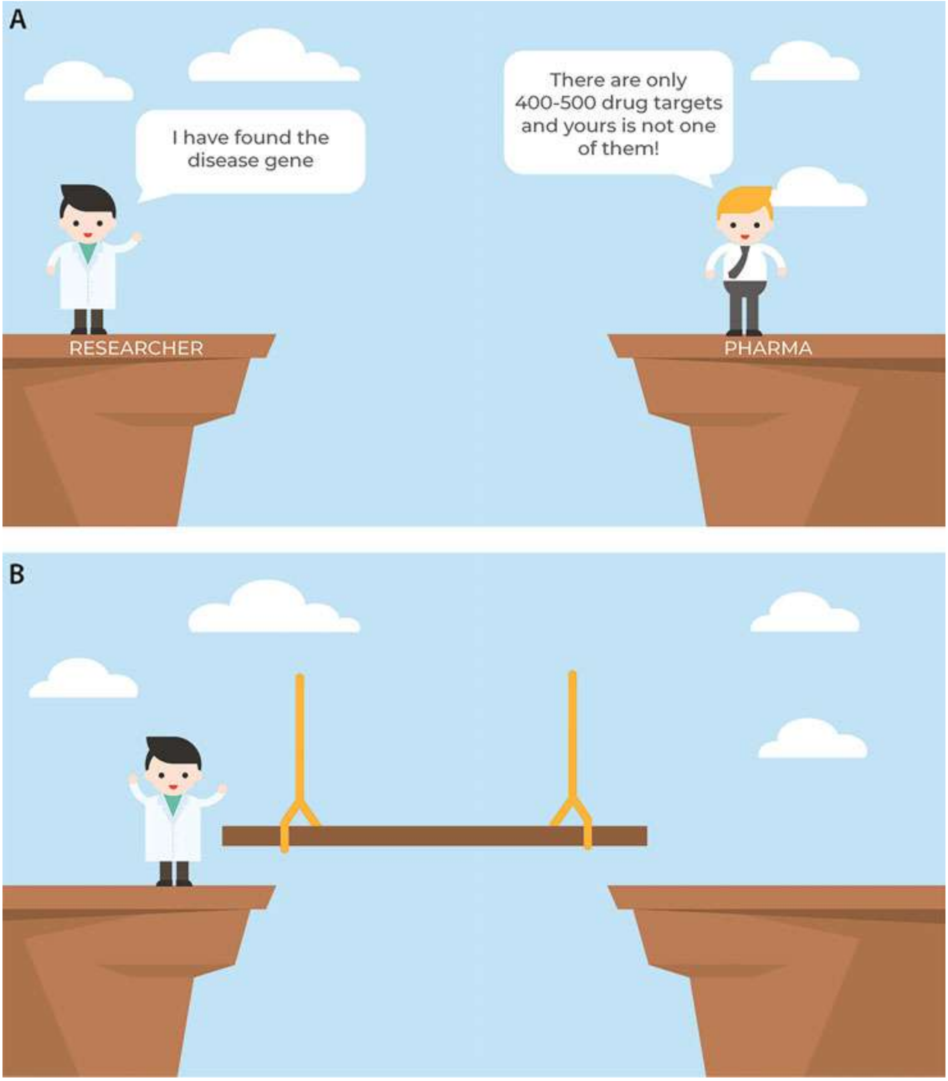

A
I have found the disease gene
There are only 400-500 drug targets and yours is not one of them!
RESEARCHER
PHARMA
B

Fig. 4.2 Illustration of the role of chemical biology in drug development. (**a**) Genomics researchers have found many new disease-associated genes whose products are not enzymes, channels, receptors, or transporters that can be addressed with the established protocols of the pharmaceutical industry. (**b**) Chemical biology acts as a bridge-builder between genomic research and drug development. Chemical probes that modulate the discovered genes help to test and validate potential therapeutic concepts (from Altmann KH et al. (2009). The State of the Art of Chemical Biology Chem Bio Chem 10: 16–29)

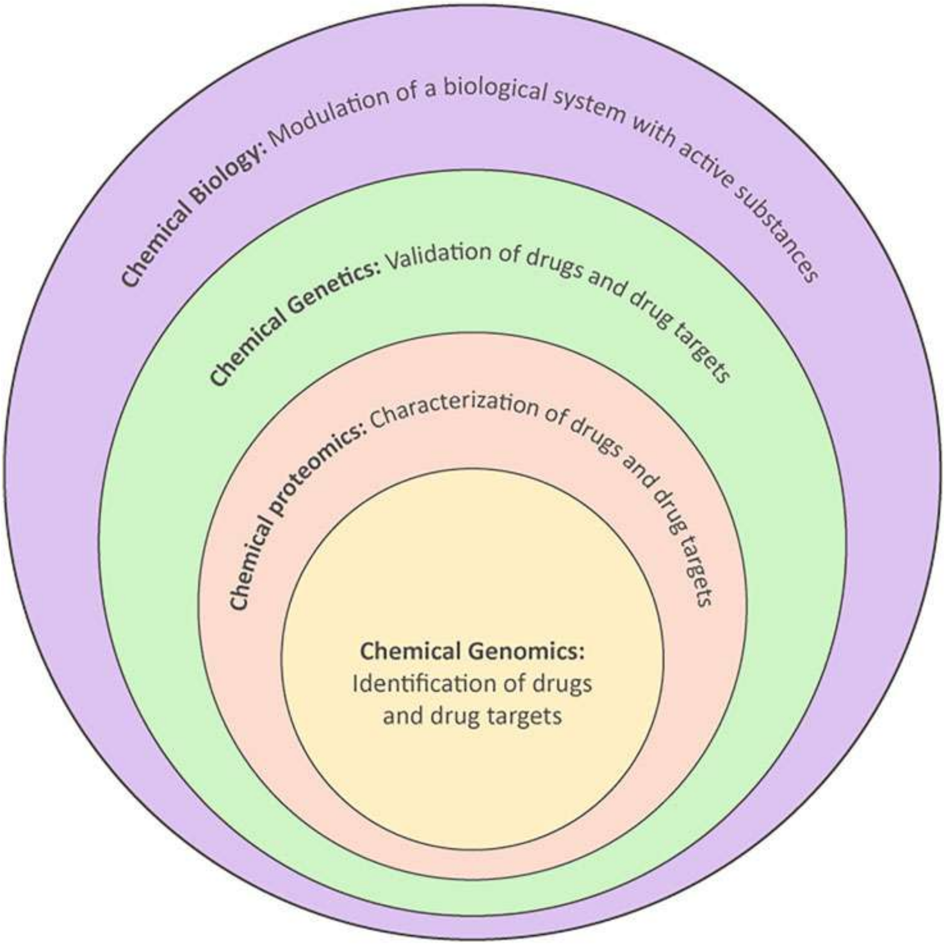

Chemical Biology: Modulation of a biological system with active substances
Chemical Genetics: Validation of drugs and drug targets
Chemical proteomics: Characterization of drugs and drug targets
Chemical Genomics: Identification of drugs and drug targets

Fig. 4.3 Chemical biology, its sub-disciplines—chemical genomics, chemical proteomics, and chemical genetics—and their tasks—identification, characterization, and validation of chemical probe and chemical probe targets

As illustrated in Fig. 4.2, genomic research has nevertheless found many new disease-associated genes whose products cannot be addressed by the established methods of drug development in the pharmaceutical industry (i.e., they are “undruggable”).

This is the reason why chemical biology plays a central role in drug discovery today. The gap between genomic research and drug discovery is being bridged by testing chemical probes to explore biological systems. The goal is the systematic

- identification,
- characterization, and
- validation of chemical probes and chemical probe targets (Fig. 4.3).

4.1 Chemical Genomics: Identification of New Drug Targets

Chemical genomics (or chemogenomics) is derived from genomics, the study of the structure of the genome (the totality of all carriers of heritable information) and the interactions between genes (the individual heritable information). As stated in the first section of Chap. 3, the sequencing of the human genome has led to the discovery of a large number of new genes. Chemical genomics aims to identify gene products that are potential drug targets. For this purpose, entire gene product or drug target families (for example, enzyme and receptor classes) are systematically tested against chemical libraries of mostly low-molecular substances, so-called probes. In other words, similar to genomics, chemical genomics is primarily concerned with the investigation of the totality of all gene products (proteins), but with a focus on their potential as drug targets and their differences from each other with respect to their binding properties and functions and less on individual genes or gene products.

Chemical genomics focuses on chemical probes (ligands), proteins (receptors), and the biological change in a trait, known as the phenotype, that results from the interaction of the two. Once the phenotype is characterized, one can associate a protein with a molecular event. In contrast to genetics, chemical genomics is able to alter the function of a protein rather than a gene. The interaction—as well as the reversibility of the interaction—can be observed in real-time. For example, the modification of a phenotype can be observed only after the addition of a specific compound and can be interrupted after its withdrawal from the medium.

There are two experimental approaches in chemical genomics: Forward (classical) chemical genomics and reverse chemical genomics (Fig. 4.4).

Forward Chemical Genomics

In forward chemical genomics, also known as classical chemical genomics, the change in the phenotype of a cell in the presence of a chemical probe is studied. For example, a phenotype change could be the arrest of tumor growth. Initially, chemical probes, also known as chemical probes, are used, and it is known which protein families they preferentially bind to. The final step is to try to identify the drug target. One tests the chemical probe that has been discovered against all known members of the known protein family.

Reverse Chemical Genomics

In reverse chemical genomics, libraries of chemical probes are tested in parallel against several members of a protein family (e.g., kinases). If one or more chemical probes are found that can modulate the function of the proteins, the probes are also tested against cells for phenotype modification. It is striking that reverse chemical genomics is almost identical to the drug target-based approach used in the pharmaceutical industry: One protein is tested against many drug candidates individually in a high-throughput manner. The only difference is that the chemical probes used in chemical genomics do not have to meet the toxicological requirements of a drug candidate and can therefore have, for example,

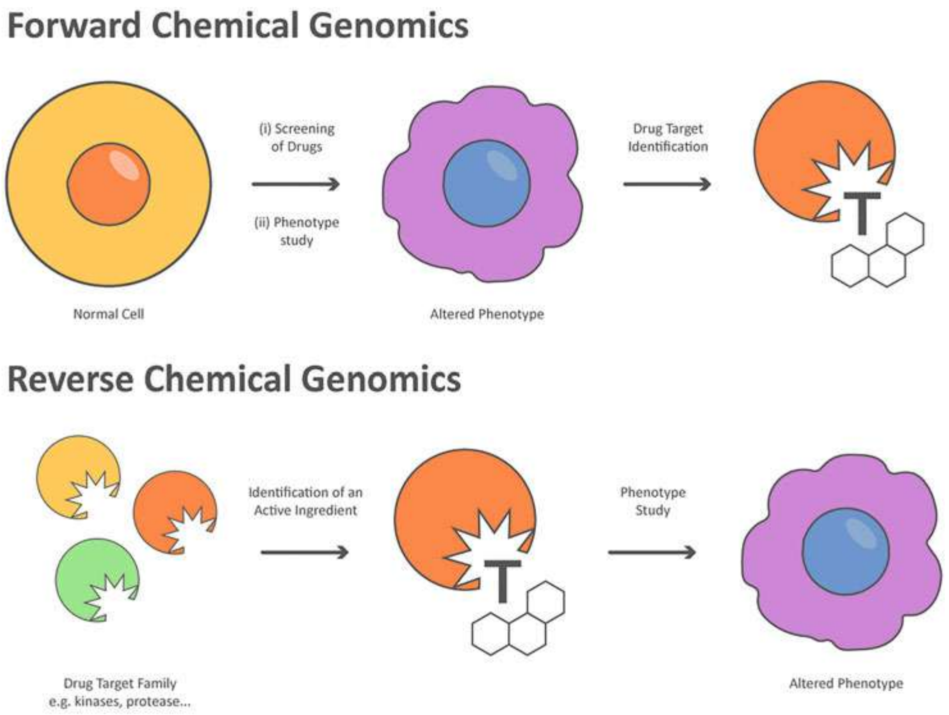

Forward Chemical Genomics
(i) Screening of Drugs
(ii) Phenotype study
Drug Target Identification
T
Normal Cell
Altered Phenotype
Reverse Chemical Genomics
Identification of an Active Ingredient
Phenotype Study
T
Drug Target Family e.g. kinases, protease...
Altered Phenotype

Fig. 4.4 The identification of new chemical probes and chemical probe targets is based on forward and reverse genetics: In forward chemical genomics, chemical probes are tested on cells in order to trigger a biological change (i.e., a change in phenotype). The target of the agent is then determined. In reverse chemical genomics, related protein families are screened against chemical libraries. If an active agent is found, it is tested on cells. In most cases, forward chemical genomics experiments go hand in hand with reverse chemical genomics experiments in order to clearly identify the drug target

chemically reactive groups. In addition, chemical genomics does not test one drug target but several (usually all known) members of a protein family in order to enable an investigation of the binding and functional differences of the proteins.

The experimental approaches of forward (classical) and reverse chemical genomics should always be performed in parallel to clearly identify the drug target.

4.2 Chemical Proteomics: Characterization of the Drug Target

Often, the clear identification of the drug target is not possible by chemical genomics, or undesired binding partners, so-called off-target effects, should be excluded. This is where chemical proteomics comes in.

Proteomics studies the proteome—the set of translated proteins at a given time under defined conditions—and aims to map the totality of all proteins and their mutual

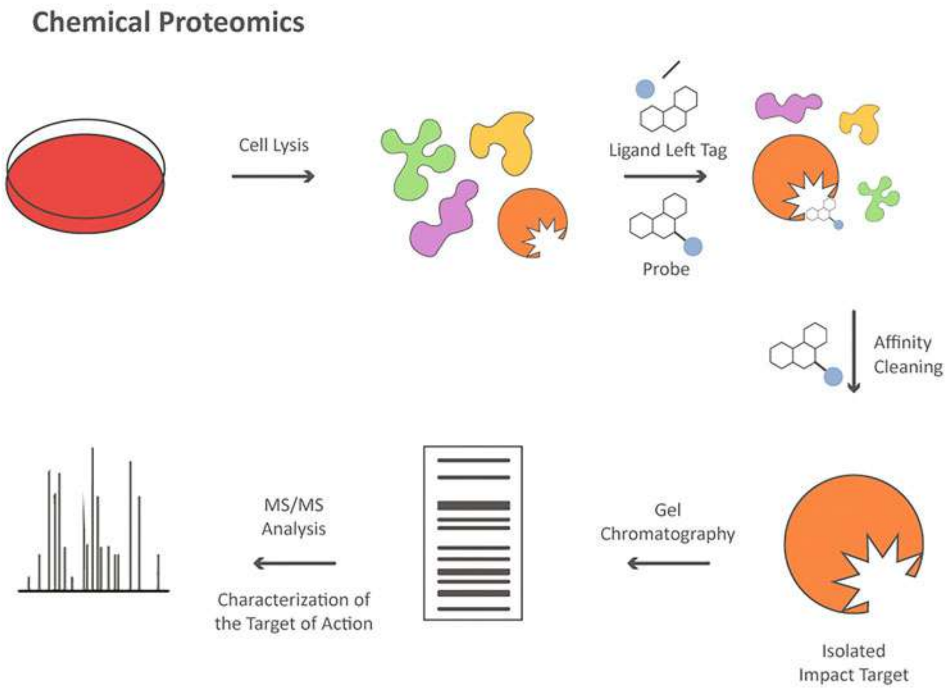


Fig. 4.5 The unambiguous identification and characterization of the ligand-protein complex is the goal of chemical proteomics. For this purpose, a ligand is chemically labeled (tagged), which allows the target protein of the ligand to be isolated via affinity chromatography. Subsequently, the target protein is examined by gel chromatography and mass spectrometry

interactions. Proteomics uses classical methods of protein biochemistry such as protein extraction and purification. However, mass spectrometry has established itself as the central method in proteomics.

In contrast, the goal of chemical proteomics is to identify and characterize the drug-drug target complex. For this purpose, a ligand is chemically linked to a tag via a linker. The generated probe is incubated with the cell lysate. Affinity purification is then performed (Fig. 4.5). For example, the tag may be a polyhistidine tag (a peptide consisting of six histidineamino acids) that binds to immobilized nickel ions via a chelate complex while one washes away the unbound proteins. Subsequently, one also washes away the bound protein from the solid phase by a solution of imidazole (the functional group of histidine). The purified target protein of the drug is then analyzed by gel chromatography and mass spectrometry. Mass spectrometry techniques also allow the amino acid sequence of the protein to be inferred, thus uniquely characterizing the protein.

4.3 Chemical Genetics: Validation of the Drug Target

Chemical genetics is the oldest subdiscipline of chemical biology and has long been synonymous with chemical biology. Therefore, it is more narrowly defined in this book than in other publications, in distinction to chemical biology.

Chemical genetics can be directly traced back to the mutagenesis experiments of genetics. As already briefly explained in Sect. 2.1 and described in Fig. 4.6, chemical genetics can be regarded as analogous to classical genetic screening. In this process, random mutations are introduced into the organism, the phenotype of these mutants is observed, and finally, the specific gene mutation (i.e., the genotype) that produced this phenotype is identified. In chemical genetics, the phenotype is not changed by the introduction of mutations, but by exposure to chemical probes.

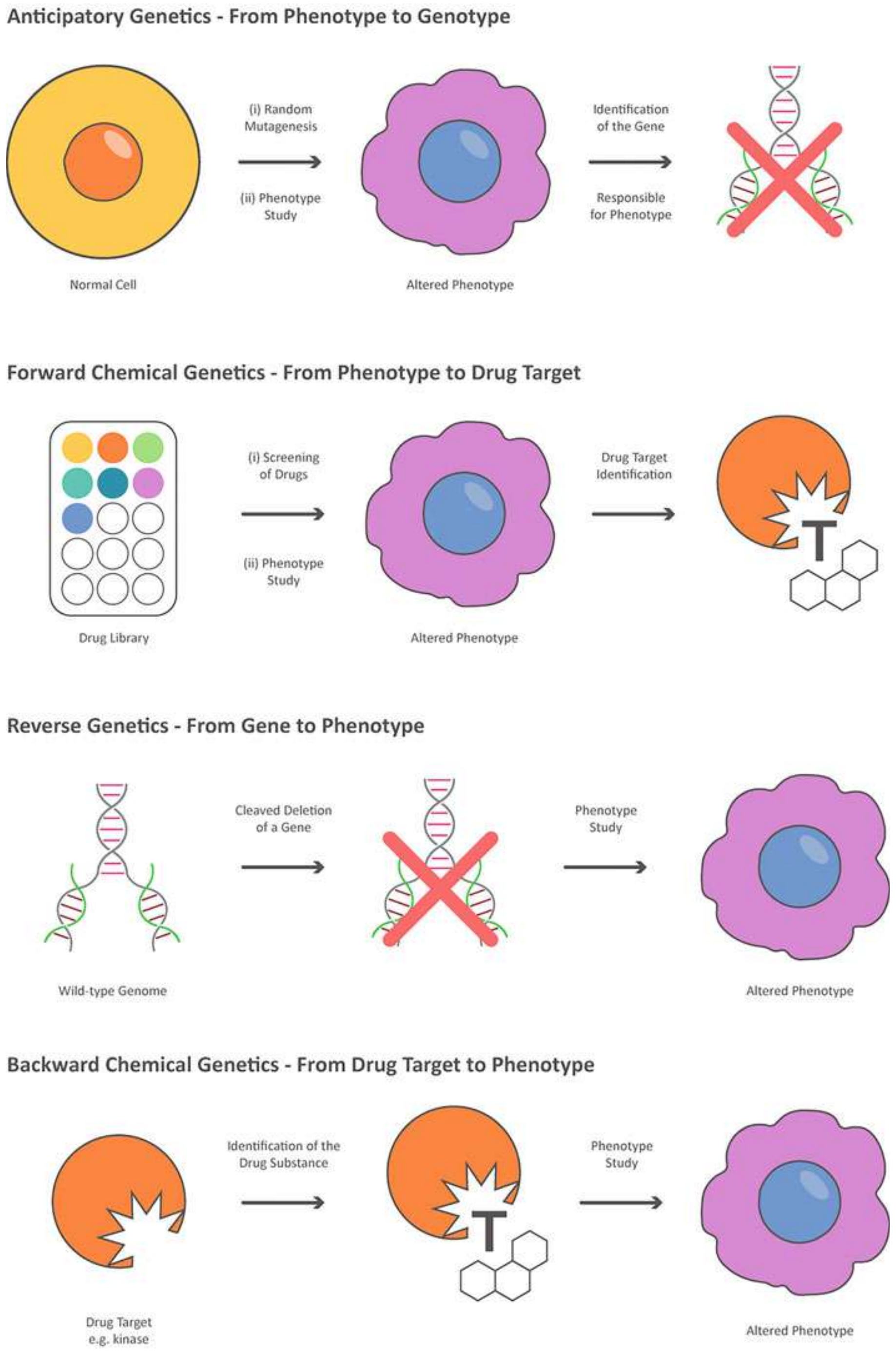
Advantages of Chemical GeneticsVersus Genetic Knock-Out Experiments

- The effect of chemical probes becomes apparent quickly.
- In most cases, the effect is reversible (due to metabolism and clearing), allowing temporal control of protein function.
- The effect is adjustable, allowing for degrees of phenotypes through different concentrations.
- The effect is conditional because the chemical probe can be introduced at any point in development. A knock-out that is lethal to the embryonic development of an organism cannot be studied for an adult organism.
- Knock-out studies cannot elucidate the role of different protein forms derived from the same gene, whereas a small molecule should in principle be able to distinguish between these functions.
- The effect can be examined by anyone who has access to the link.

Disadvantages of Chemical GeneticsVersus Genetic Knock-Out Experiments

- A compound must first be identified that has an effect.
- The identification of the biological target of the ligand can be very complex.

As in chemical genomics, experiments are done in a forward and reverse manner. Phenotypic screening (forward) of chemical libraries is used to find an agent that can alter the phenotype. Reverse chemical genetics plays a special role in validating drug targets in experimental disease models (i.e., in target validation): Does manipulate this protein have the desired clinical effect?



Anticipatory Genetics - From Phenotype to Genotype

Forward Chemical Genetics - From Phenotype to Drug Target

Reverse Genetics - From Gene to Phenotype

Backward Chemical Genetics - From Drug Target to Phenotype

Fig. 4.6 Chemical genetics is used to validate a protein for its suitability as a drug target. Chemical genetics can be traced back to mutagenesis experiments in genetics. Instead of introducing mutations into the genome, chemical probes are introduced and changes in phenotype are observed. Analogous

It should therefore be explicitly stated: The experimental design in chemical genetics is no different from that in chemical genomics. In chemical genetics, the focus is on validating the drug target. In order to achieve the desired clinical effect in humans, it must first be shown that this effect is detectable in cell experiments. As described in Sect. 3.2, the failure to predict the clinical efficacy of a compound is the most important reason for the failure of drug development.

4.4 Chemical Biology

In addition to the subfields of chemical genomics (drug identification), chemical proteomics (drug target characterization), and chemical genetics (drug target validation), an increasing number of approaches are playing a role in drug development that makes it possible to manipulate previously unaddressable drug targets with chemical probes. In the pharmaceutical industry, these strategies are now grouped under the generic term *chemical biology*.

In classical drug discovery and pharmacology, there is *a drug* that binds to *a* drug target that alters the phenotype or shows a clinical effect (Fig. 4.7). As we have learned, only about 3% of human genes code for drug targets currently in clinical use. Nevertheless, how can we manipulate a protein that belongs to the other 97%? Approaches such as antibody-drug conjugates (ADC), the combination of a monoclonal antibody with a small molecule ligand for targeted inhibition of tumor cells, or proteolysis targeting chimera (PROTAC) are new techniques that allow proteins outside the current spectrum of activity to be targeted. In terms of drug development, chemical biology has established itself as the generic term for these processes and approaches.

Summary

Genomic research has provided access to many new potential drug targets. The task of chemical biology and its subdisciplines is to test these targets as the basis for new therapies. While chemical genomics systematically identifies ligands for new drug targets, chemical proteomics characterizes the binding of the ligand to the drug target, and chemical genetics validates the gene product found as a drug target. The generic term chemical biology covers all approaches to address drug targets that could not previously be modulated by small-molecule ligands and biologics.

Fig. 4.6 (continued) to chemical genomics, the two experimental approaches of forward and reverse chemical genetics exist. Chemical genomics and chemical genetics do not differ in their experimental implementation. They differ only in their goals of identifying and validating the drug target

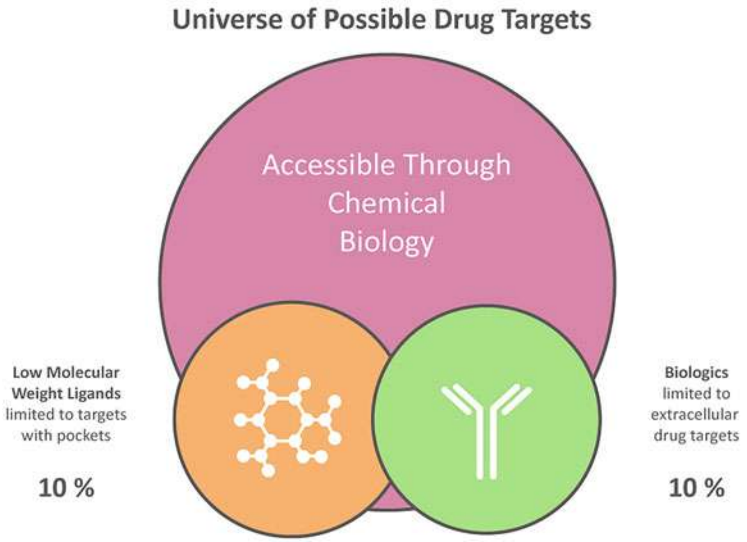

Universe of Possible Drug Targets
Accessible Through Chemical Biology
Low Molecular Weight Ligands limited to targets with pockets
10 %
Biologics limited to extracellular drug targets
10 %

Fig. 4.7 The problem of non-addressable drug targets. If we look at the proteins encoded in the genome, we notice that only 10% of them either have a deep, hydrophobic pocket in which a small-molecule ligand can bind or else occur extracellularly, subject to binding with a biologic. It is assumed that 80% of all possible drug targets cannot be addressed with small-molecule ligands or biologics. In contrast, chemical biology holds the promise of being able to modulate previously unaddressable proteins by means of new concepts of action. In pharmacology, *a* drug manipulates *a* drug target, which triggers an altered phenotype. The basis of new active concepts in chemical biology are combinations of chemical probes. Previously unmanipulable proteins are to be modulated by the network

Further Reading

Altmann KH et al (2009) The state of the art of chemical biology. Chem Bio Chem 10:16–29

Bredel M, Jacoby E (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. Nat Rev Genet 5:262–275

Santos R et al (2017) A comprehensive map of molecular drug targets. Nat Rev Drug Discov 16:19–34

Part II

DNA

DNA: Blueprint of the Proteins

5

Deoxyribonucleic acid (DNA) is a molecule consisting of two chains that wrap around each other to form a double helix. It contains the genetic instructions for the development, function, growth, and reproduction of all known organisms and many viruses (Fig. 5.1). DNA and ribonucleic acid (RNA) are nucleic acids. Along with proteins, lipids, and complex carbohydrates (polysaccharides), nucleic acids are one of the four main types of macromolecules essential to all known forms of life.

5.1 Function and Structure of DNA

The two strands of DNA are also called polynucleotides because they consist of simpler monomer units called nucleotides. Each nucleotide consists of one of four nitrogen-containing nucleobases (cytosine [C], guanine [G], adenine [A], or thymine [T]), a sugar called deoxyribose, and a phosphate group. The nucleotides are linked together in a chain by covalent bonds between the sugar of one nucleotide and the phosphate residue of the next nucleotide, resulting in an alternating sugar-phosphate backbone. The nitrogen-containing bases of the two separate polynucleotide strands are bonded together with hydrogen bonds to form double-stranded DNA according to the base-pairing rules (A with T and C with G). The complementary nitrogenous bases are divided into two groups, pyrimidines, and purines. In DNA, the pyrimidines are thymine and cytosine; the purines are adenine and guanine (Fig. 5.2).

Both strands of double-stranded DNA store the same biological information. This information is replicated as soon as the two strands separate. A large proportion of DNA (more than 98% for humans) is non-coding, which means that these sections do not serve as patterns for protein sequences (Figs. 5.3, 5.4, and 5.5). The two DNA strands run in opposite directions and are therefore antiparallel. One of four types of nucleobases (bases

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_5

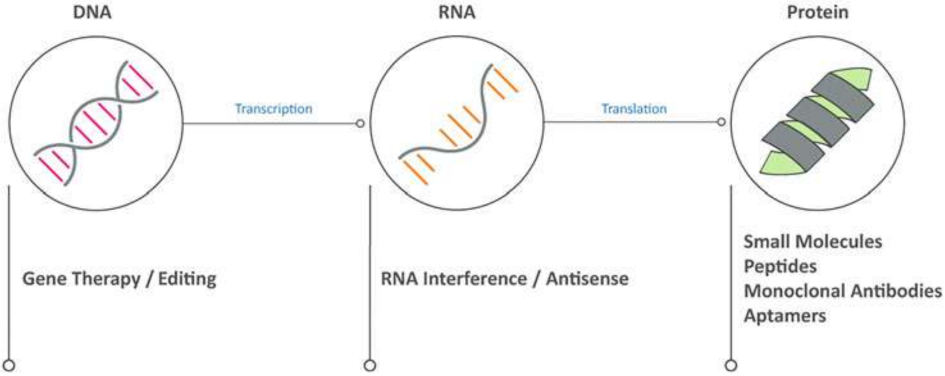


Fig. 5.1 The central dogma of molecular biology describes the transfer of information from DNA via RNA to the functional carrier proteins, which is determined by the order (sequence) of the respective monomers (nucleotides in the case of DNA and RNA; amino acids in the case of proteins). The function of the three biopolymers can be influenced by chemical compounds or by biochemical processes for chemical-biological experiments

for short) is attached to each sugar residue. It is the sequence of these four nucleobases along the backbone that encodes the genetic information. RNA strands are generated using DNA strands as a template in a process known as transcription, in which DNA bases are exchanged for their corresponding bases, except in the case of thymine (T), which is replaced in RNA by uracil (U). Under the genetic code, these RNA strands specify the sequence of amino acids within proteins in a process called translation.

In eukaryotic cells, DNA is organized into long structures called chromosomes (Fig. 5.3). Prior to typical cell division, these chromosomes are duplicated during DNA replication (Fig. 5.4), providing a complete set of chromosomes for each daughter cell. Eukaryotic organisms (animals, plants, fungi, and protists) store most of their DNA in the nucleus as nuclear DNA and some in mitochondria as mitochondrial DNA or in chloroplasts as chloroplast DNA. In contrast, prokaryotes (bacteria and archaea) store their DNA only in the cytoplasm in circular chromosomes. In eukaryotic chromosomes, chromatin proteins, such as histones compact and organize DNA. These compaction structures control the interactions between DNA and other proteins and help to control which parts of the DNA are transcribed.

5.2 Oligonucleotide Synthesis

Oligonucleotides are synthesized in the cell by enzymes called polymerases. However, polymerases require certain start sequences that later remain in the oligonucleotide. In order to obtain shorter oligonucleotides of any base sequence, phosphoramidite synthesis has become established.

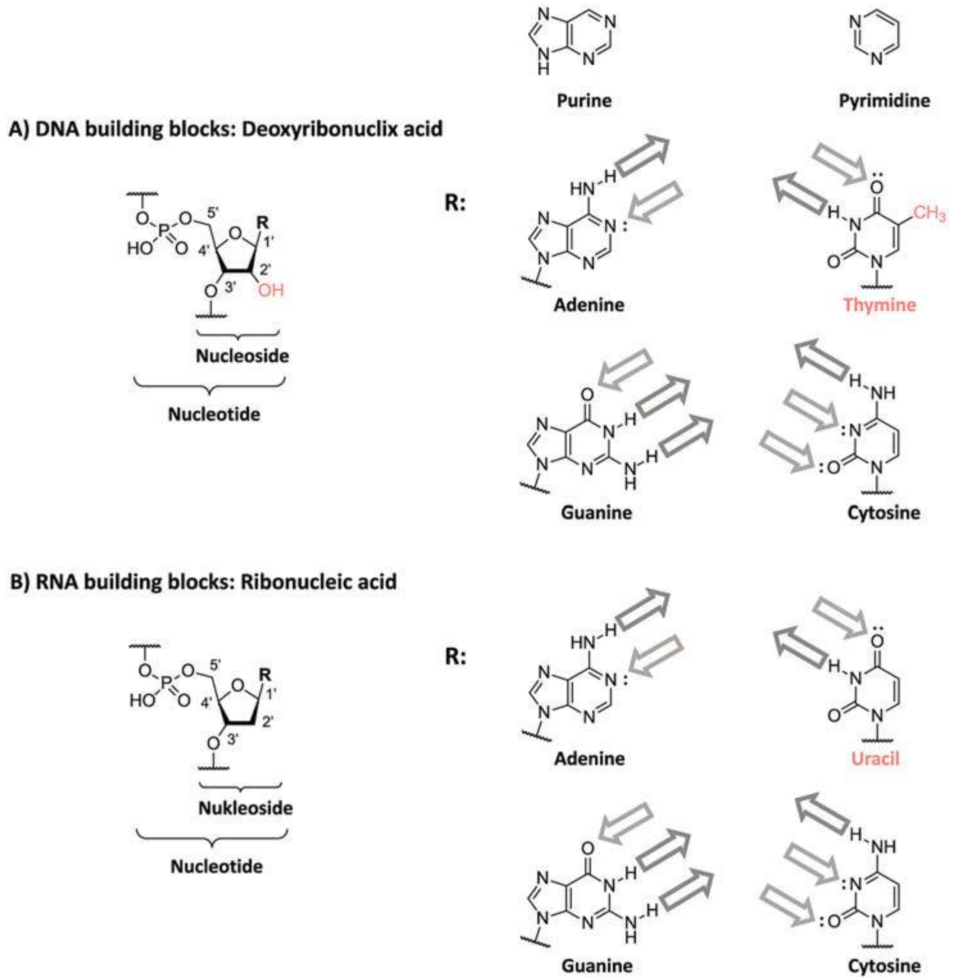


Fig. 5.2 Building blocks of DNA: Bases and sugar backbone

For this purpose, as shown in Fig. 5.6, protected phosphoramidites are used as nucleotide analogs for the synthesis of oligonucleotides. The 5-hydroxy group is protected with a dimethoxytrityl (DMT) group. In the case of ribonucleic acid (RNA), the 2-hydroxy group is orthogonally protected with a *tert-butyldimethylsilyl* (TBDMS) or tri-iso-propylsilyloxymethyl (TOM) group. The 3-hydroxy group is reacted with a 2-cyanoethyl-N,N,N′,N″-tetraisopropylphosphorodiamidite to give the phosphoramidite. As also shown in Fig. 5.6, except for thymine and uracil, the bases are protected: Cytosine is acetyl protected, Adenine is benzyl protected and Guanine is isopropyl protected.

Oligonucleotide synthesis takes place in the solid phase. For this purpose, a nucleotide building block is immobilized at the 3′-terminus. By stepwise addition of nucleotide

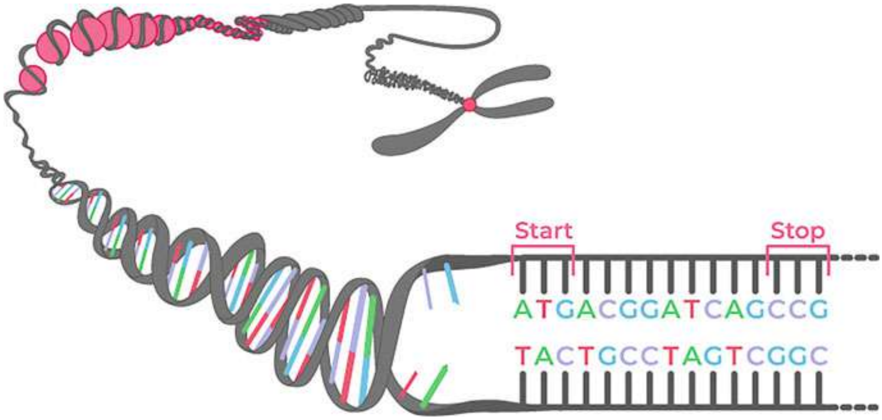


Fig. 5.3 Structure of a chromosome and model representation of the genes on it

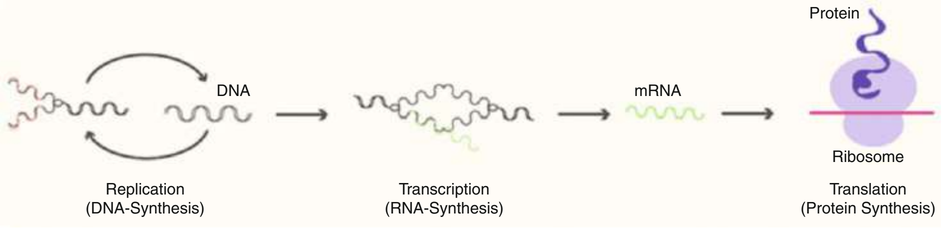


Fig. 5.4 Based on the central dogma of molecular biology, the information transfer from protein-coding DNA via derived RNA synthesis (transcription) to protein biosynthesis (translation) is shown below

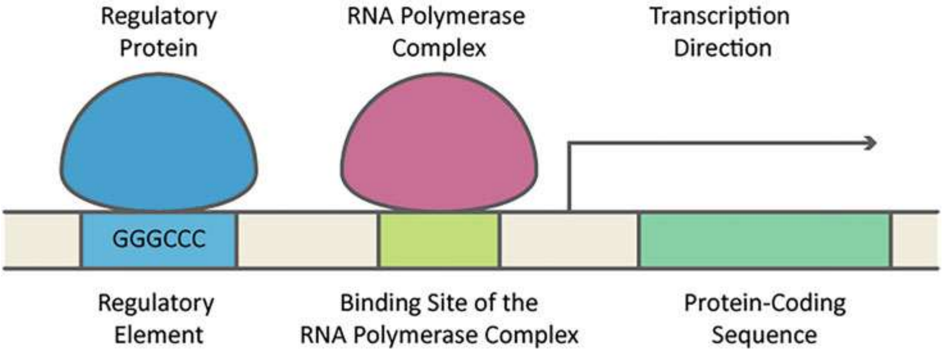


Fig. 5.5 General structure of a gene: Starting with a regulatory binding site for a transcription factor followed by a binding site for the RNA polymerase, which after binding reads the protein-coding sequence and synthesizes it into an mRNA in the process

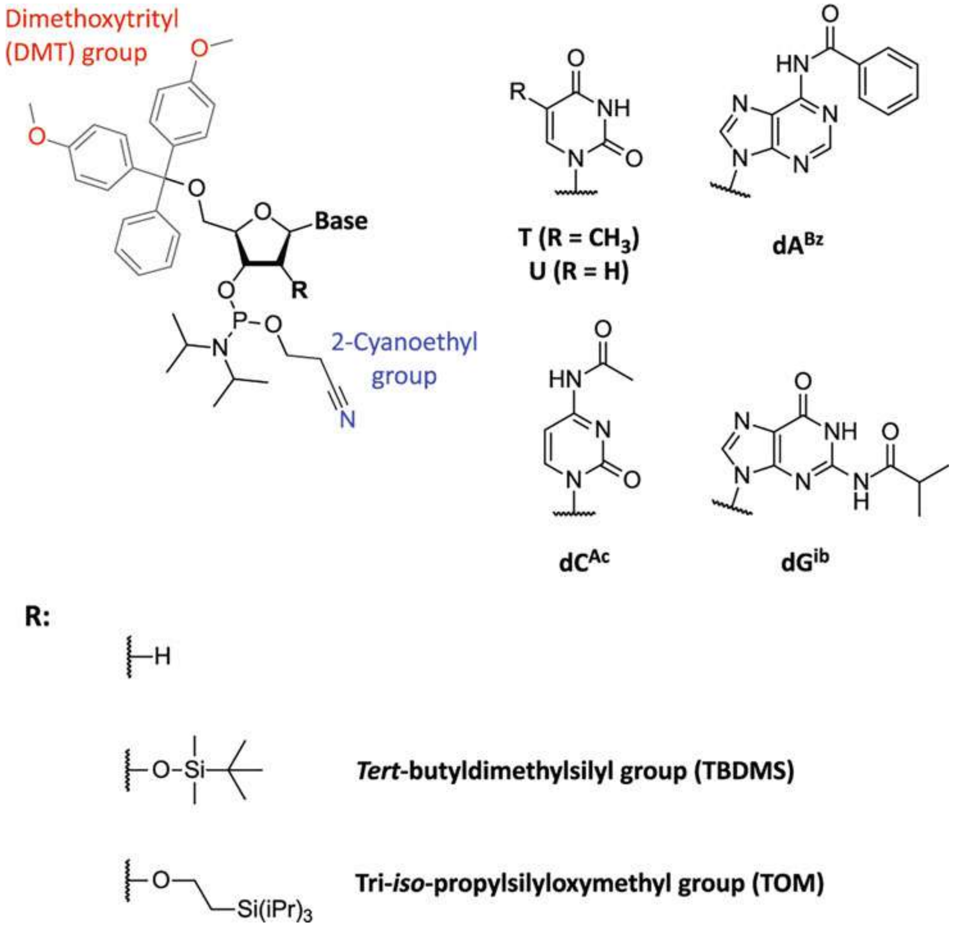


Fig. 5.6 Overview of the synthesis building blocks in oligonucleotide synthesis

residues to the 5′-terminus of the nucleotide building blocks, the chain grows until the desired sequence is assembled. Each addition is called a synthesis cycle (Fig. 5.7) and consists of four chemical reactions:

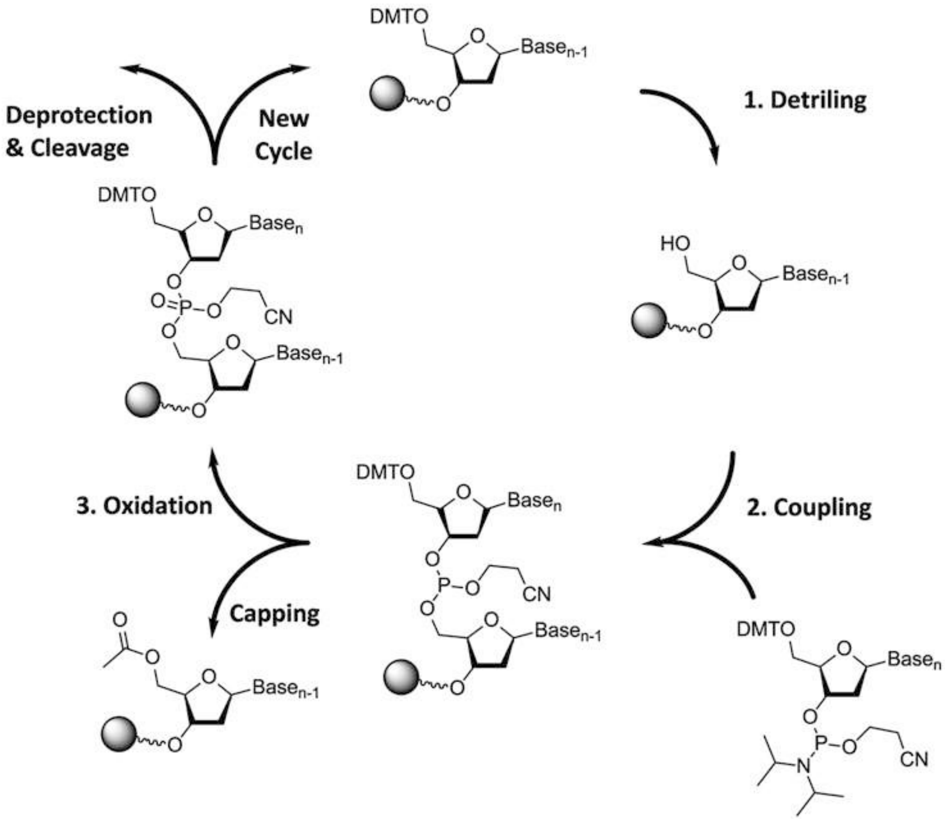


Fig. 5.7 Synthesis cycle of oligonucleotide synthesis

Overview

Step 1: Unlock (Detritylation)

The DMT group is cleaved with a solution of an acid such as 2% trichloroacetic acid (TCA) or 3% dichloroacetic acid (DCA) in an inert solvent such as dichloromethane or toluene. The cleaved DMT cation is washed away; the step results in the oligonucleotide precursor bound to solid support bearing a free 5-′-terminal hydroxy group.

Step 2: Coupling

The nucleoside phosphoramidite is dissolved in acetonitrile and activated with an acidic azole catalyst such as 1H-tetrazole or 5-ethylthio-1H-tetrazole before addition to the solid phase. Meanwhile, mixing is usually done just before the addition to the solid phase. The activated phosphoramidite is added in excess (1.5 to 20) over the

(continued)

support-bound material. The 5′-hydroxy group of the bound nucleotide building block reacts with the activated phosphoramidite moiety to form a phosphite-triester bond. The coupling of 2′-deoxynucleoside phosphoramidites is very rapid and complete after approximately 20 s. The reaction is very sensitive to water as an alternative nucleophile. Therefore, anhydrous acetonitrile is used. After completion of the coupling, unbound reagents and by-products are removed by washing.

Step 3: Capping

Capping of the uncoupled 5′-hydroxy group bound to the solid phase is performed by the addition of acetic anhydride and serves the purpose of preventing chain formation of unreacted 5′-OH groups in the next synthetic cycle.

Step 4: Oxidation

The newly formed phosphitetriester bond is not stable. Treatment of the carrier-bound material with iodine and water in the presence of a weak base such as pyridine oxidizes the phosphitetriester to a stable phosphate triester.

These steps can be repeated as required until the desired sequence is obtained. In the last step, cleavage takes place depending on the selected linker to which the first nucleotide building block was attached. In most cases nowadays, non-nucleoside linkers are used, which are cleaved basically using ammonia in methanol.

5.3 DNA Sequencing

DNA sequencing involves determining the nucleic acid sequence—the order of nucleotides in DNA. It includes any method or technology that determines the order of the four bases: adenine, guanine, cytosine, and thymine. The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery.

5.3.1 Polymerase Chain Reaction—PCR

Polymerase chain reaction (PCR) is a method for the in vitro amplification of deoxyribonucleic acid (DNA). The basis of PCR is an enzyme, the thermostable DNA polymerase, which can incorporate the DNA building blocks. Analogous to the iterative synthesis cycle (Fig. 5.7), amplification occurs in cycles that can be repeated at will, all of which have the previous product as their output. This allows for exponential amplification.

PCR is the central method of molecular biology. Ultimately, all DNA sequencing is based directly and indirectly on this technique.

The amplification of a DNA strand requires the following components:

- The DNA template with the target region to be amplified

- A heat-resistant DNA polymerase, which remains intact during the denaturation step at 95 °C
- Two DNA primers complementary to the 3′-ends of each sense and antisense strand of the target DNA, since the DNA polymerase can bind to only one double-stranded region and incorporate new DNA building blocks from there
- Deoxy-nucleotide triphosphates (dNTP: dATP, dCTP, dGTP, dTTP)—the designation of the nucleotides containing triphosphate groups—from which the DNA polymerase builds the new DNA strand
- Buffer solution with magnesium ions Mg^{2+} as a cofactor of thermostable DNA polymerase

Procedure

The PCR process usually comprises 20 to 50 cycles. These are usually performed in a so-called thermocycler but can also be carried out in three water baths at different temperatures. A single cycle consists of three parts (Fig. 5.8):

Overview

1. **Denaturation (Melting)**

   Denaturation, or the breaking of the hydrogen bonds of the double-stranded DNA template into two single strands, is the first step in the amplification cycle. The reaction chamber is heated to 94–98 °C for 20–30 seconds.
2. **Primer Annealing**

   In the next step, the reaction temperature is lowered to 50–65 °C for 20–40 seconds, which causes the primers to bind to each of the single-stranded DNA templates. The reaction mixture typically contains two different primers: One for each of the two single-stranded complements containing the target region. The primers are themselves single-stranded sequences but much shorter than the length of the target region and complement only very short sequences at the 3′-end of each strand. It is important to determine an appropriate temperature for annealing, as efficiency and specificity are strongly influenced by temperature. This temperature must be low enough to allow annealing of the primer to the strand but high enough so that the annealing is specific, i.e., the primer should bind only to a perfectly complementary part of the strand and nowhere else. If the temperature is too low, the primer will not bind perfectly. If the temperature is too high, the primer may not bind at all. A typical annealing temperature is about 3–5 °C below the melting temperature (T_m) of the primers used. Stable hydrogen bonds between complementary bases are only formed if the primer sequence matches the template sequence very closely.
3. **Elongation (amplification)**

   In the final step, the DNA polymerase binds to the primer-template hybrid and begins to form the complementary DNA strand. The temperature in this step depends on the DNA polymerase used: The optimal activity temperature for the thermostable DNA

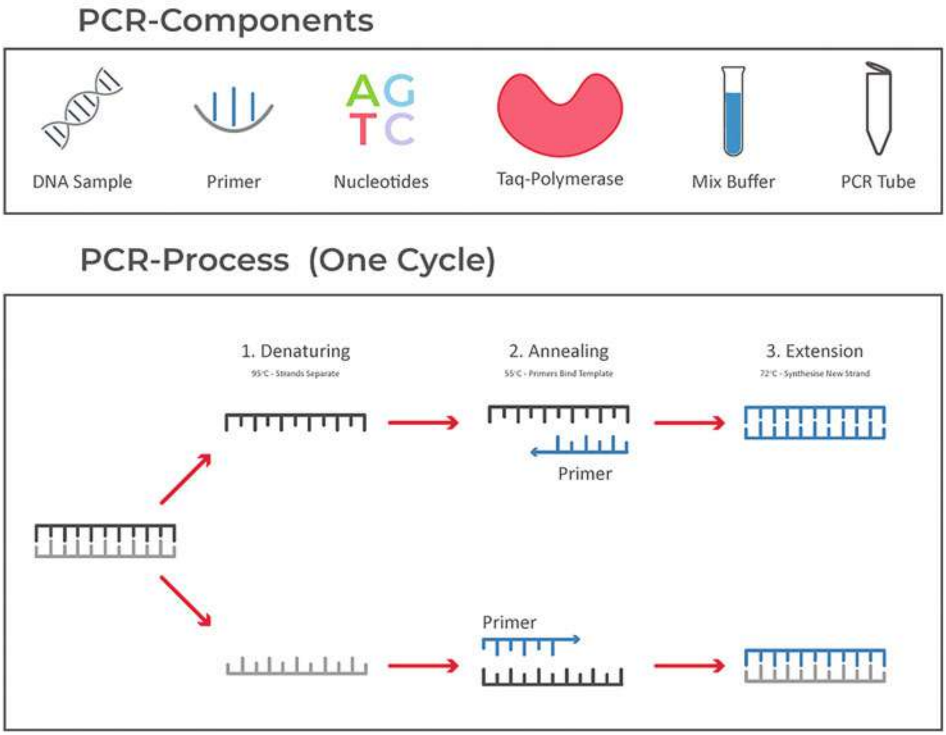


Fig. 5.8 Representation of polymerase chain reaction (PCR). (**a**) The components for PCR: DNA sample, primers, dNTPs, Taq polymerase, buffer with magnesium ions, and a reaction vessel; (**b**) Sequence of the PCR: 1. denaturation of the double-strand at 95 °C, 2. annealing of the primers at 55 °C, 3. extension at 72 °C

polymerase Taq polymerase (*Thermus aquaticus*) is approximately 75–80 °C. The DNA polymerase synthesizes a new DNA strand by adding free dNTPs from the reaction mixture that is complementary to the template in the 5′-to-3′ direction, condensing the 5′-phosphate group of the dNTPs. The exact time required for extension depends on both the DNA polymerase used and the length of the DNA target region to be amplified. As a rule of thumb, most DNA polymerases polymerize a thousand bases per minute at their optimal temperature. Under optimal conditions (i.e., when there are no limitations due to limiting substrates or reagents), each extension step doubles the number of DNA target sequences. With each successive cycle, the original template strands plus any newly generated strands become additional template strands for the next round of extension, resulting in exponential amplification of the specific DNA target region.

The sub-steps of denaturation, annealing, and elongation form a single cycle. Several cycles are required to amplify the DNA as much as desired. Therefore, the formula for calculating the number of DNA copies formed after a given number of cycles is 2^n, where n is the number of cycles. Thus, one set of reactions for 30 cycles results in 2^{30} or 1,073,741,824 copies of the original double-stranded target DNA.

5.3.2 Sanger Termination Sequencing

Sequencing is the process of determining the sequence of the four building blocks adenine, cytosine, guanine, and thymine. Based on the sequence, the coded amino acid and thus the protein can be inferred. The first and for a long time the only method for DNA sequencing was Sanger sequencing, developed in 1977 and named after its inventor, Frederick Sanger. Sanger sequencing is a modification of PCR with the difference that chain-terminating didesoxynucleotides (ddNTPs) are added to the dNTPs. Since the turn of the millennium, more economical methods for sequencing whole genomes (next generation sequencing) have been developed. However, Sanger sequencing is still used today for sequencing short DNA strands of up to 500 nucleotides because of its high accuracy.

As shown in Fig. 5.9, the classical Sanger chain termination method, similar to PCR, requires

- a single-stranded DNA template whose sequence you want to know,
- DNA primers that bind to the template,
- DNA polymerase,
- deoxyribonucleotide triphosphates (dNTPs),
- dideoxyribonucleotide triphosphates (ddNTPs), which terminate chain elongation catalyzed by DNA polymerase.

As mentioned, Sanger sequencing relies on dideoxyribonucleotide triphosphates (ddNTPs). These chain-terminating nucleotides lack a 3′-OH group, which is required for the formation of a phosphodiester bond between two nucleotides, resulting in DNA polymerase termination of DNA elongation when a modified ddNTP is incorporated. The concentration of the didesoxynucleotides (ddNTPs) should be approximately 100-fold higher than that of the corresponding deoxynucleotide (e.g., 0.5 mM ddTTP to 0.005 mM dTTP) to produce sufficient fragments while still transcribing the complete sequence. The ddNTPs are usually radioactively or fluorescently labeled according to their base for their detection. In the subsequent gel electrophoresis, the PCR products of different lengths are separated. Based on the specific fluorescence signal for a base, the corresponding base can be inferred for each position.

Sanger sequencing marks an important milestone in genomic research. However, only short pieces of DNA can ever be sequenced using this method because a) the lengths of the gels used in electrophoresis are limited and b) from an economic point of view, enormous

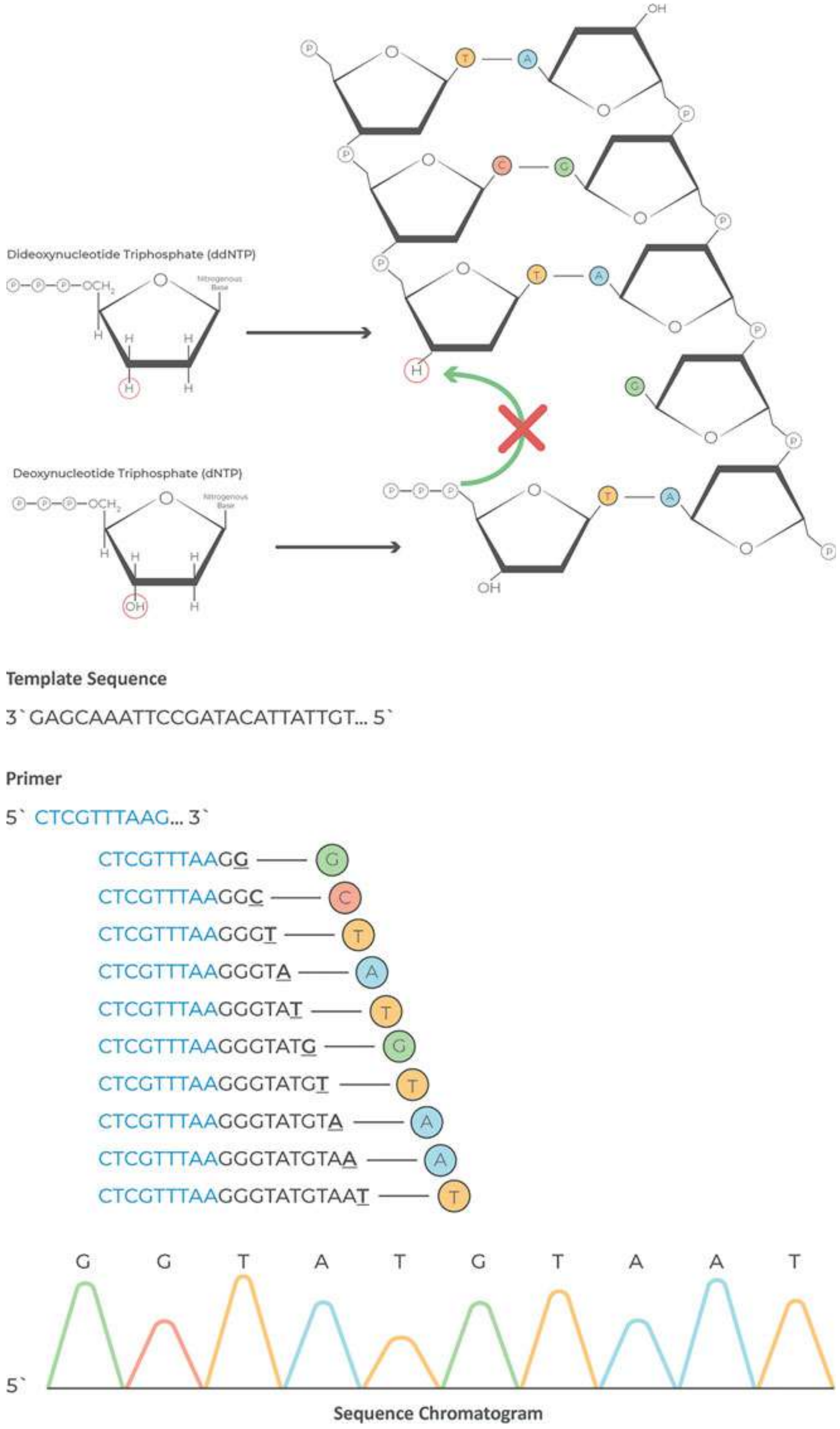


Fig. 5.9 Sanger sequencing with the truncation method using ddNTPs

amounts of DNA are required due to chain termination. The first sequencing of a human genome was basically based on Sanger sequencing. The cost of that was more than one billion euro. And this was only possible at the time because large strands of DNA could be specifically cut into smaller strands using restriction endonucleases and it was known where the restriction endonucleases specifically cut the DNA. In addition, it was only the greatly increased computing power of computers from the end of the 1990s that made it possible to reassemble these snippets from a large number of Sanger sequencing experiments.

5.3.3 Next Generation Sequencing

In contrast to classical Sanger sequencing, a number of economically more efficient whole-genome sequencing (WGS) methods have become established. The cost of sequencing has dropped from over a billion euros with Sanger sequencing to a few hundred euros. These new methods fall under the term next generation sequencing (NGS). In the meantime, there are a large number of different NGS methods on the market, and it is easy to lose track of them all. What is more important to understand is what all of these methods have in common technically: Sequencing by synthesis.

As in Sect. 5.3.1, the new genome sequencing methods use restriction endonucleases to specifically cut DNA into small pieces. With the now-even-faster computers, the sequencing products can be assembled even more efficiently into the whole genome. The key innovation of all these methods has been that the amount of DNA required has been drastically reduced by avoiding irreversible termination of chain extension. Figure 5.10 shows the DNA building blocks of sequencing according to Solexa. The fluorophore is specific to the base in question and, like the 3′-OH group, is protected with an azide. This means that after the incorporation of the base, a fluorescent signal can be used to detect the base at that site. The next step is the chemical attack on the two azides. As a result, the fluorophore is cleaved off and the 3′-OH group is present deprotected. In the next step, DNA polymerase couples the next building block based on the complementary DNA template. The base is again detected based on the fluorophore. Subsequently, the fluorophore is cleaved off again so that it does not interfere with the next step. In addition, the 3′-OH group is deprotected in order to be able to couple with the next DNA building block.

In contrast to the Sanger method, only one DNA strand is required and not several DNA strands that continuously break off after each base. This sequencing by synthesis saves enormous amounts of DNA.

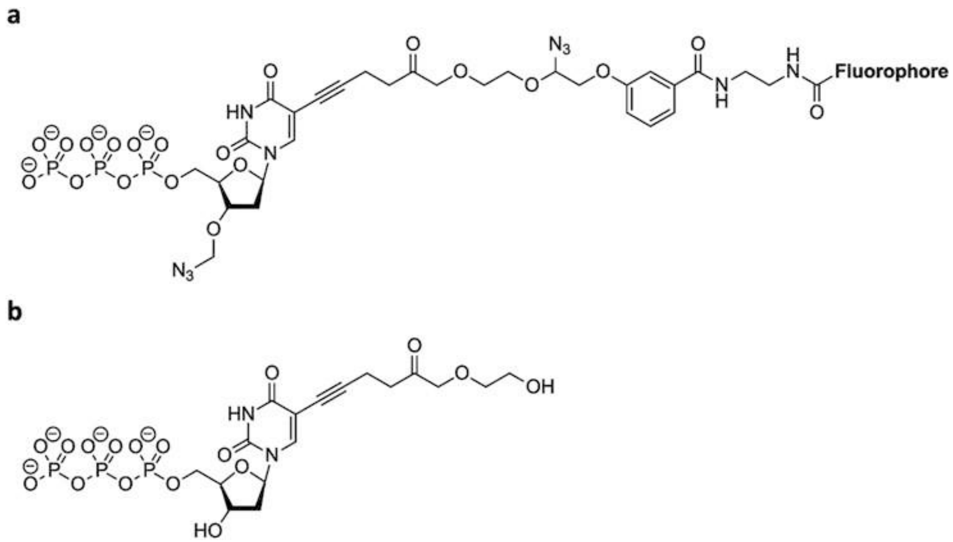


Fig. 5.10 Example of sequencing by synthesis: (**a**) The building block with the base-specific fluorophore and the protected 3′-OH group. (**b**) With the cleavage of the fluorophore, the 3′-OH group is simultaneously released to participate in the next synthesis cycle

5.3.4 DNA Microarray

Another inexpensive method for (non-complete) DNA sequencing is the DNA microarray. A DNA microarray (also commonly known as a DNA chip or biochip) is a collection of microscopic DNA spots adhered to a solid surface.

Originally, this method was used to detect the expression of a gene by detecting the mRNA formed. For this purpose, a few picomoles (10^{-12} mol) of the DNA sequence to be tested, which is provided with a fluorescent probe, are immobilized on a carrier (usually glass). In the next step, the mRNA to be tested is added. If one of the mRNAs binds to the immobilized probes, the fluorescence signal is quenched, i.e., suppressed. Since it is precisely defined which probe was applied to which position on the chip, it can be concluded which mRNA was bound or which gene was expressed.

Meanwhile, DNA microarrays can also be used for sequencing (Fig. 5.11). Due to the large sequencing projects, we know that DNA consists of approximately 3.2 billion base pairs. The individual genomes differ approximately 335 million positions, so-called single nucleotide polymorphisms(SNPs). Of these, about 15 million occur in different populations worldwide with a frequency of more than 1%. Instead of cost-intensive sequencing of the entire genome, only the most frequent differences (single nucleotide polymorphisms) are tested with the chip.

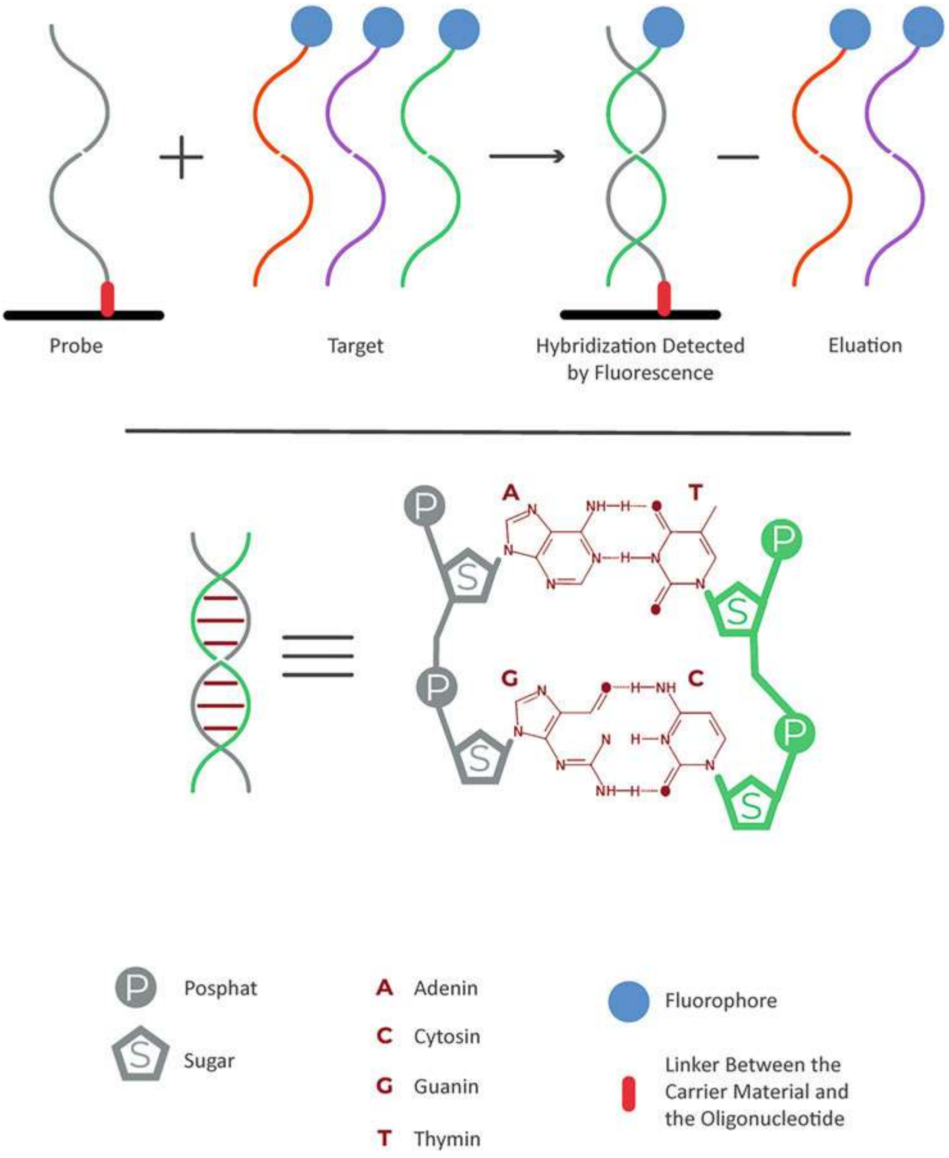

Probe
Target
Hybridization Detected by Fluorescence
Eluation
A
T
G
C
P
S
NH—H
N—H—N
H—N
P Posphat
S Sugar
A Adenin
C Cytosin
G Guanin
T Thymin
Fluorophore
Linker Between the Carrier Material and the Oligonucleotide

Fig. 5.11 Example of genotyping using a DNA microarray

For this purpose, allele-specific oligonucleotide (ASO) probes are often selected based on sequencing a representative group of individuals: Positions found to vary with a certain frequency in the group are used as the basis for probes. In this context, SNP chips are generally described by the number of SNP positions they test. Unlike a gene expression test, two probes must be used for each SNP position to detect both alleles. If only one probe were used, homozygosity could not be distinguished from heterozygosity of the tested allele.

Summary

The genome consists of long chains of the deoxynucleic acid building blocks—adenine, thymine, guanine, and cytosine—which are wound onto proteins called histones to form the individual chromosomes. The sequence of DNA letters codes for proteins. Short DNA and RNA oligonucleotides can be rapidly synthesized in the solid phase. Of greater interest has long been the elucidation of the sequence (i.e., sequencing). Necessary techniques and technologies include polymerase chain reaction, Sanger termination sequencing, next generation sequencing by synthesis, and DNA microarrays.

6 Genomic Variation

All humans have almost the same genome with an almost identical DNA sequence of more than 3.2 billion base pairs (i.e. 6.4 billion DNA letters). Only slight differences exist between individuals, making each of us unique. These differences, called genetic variants, occur at specific locations within the DNA. The sequence of DNA building blocks is a code. This code consists of the four building blocks adenine, thymine, cytosine, and guanine, abbreviated by the letters A, T, C, and G. A genetic variant occurs at a location within the DNA where this code differs between people.

The simplest form of genetic variation is a difference in the nucleotide base present. These types of variants are called single nucleotide polymorphisms (SNP). Out of 3.2 billion base pairs, people differ in about 335 million positions. Of these, only 15 million occur more frequently than 1% across multiple populations worldwide. Figure 6.1 shows the two copies of a DNA molecule from the same region of two different individuals. Note that the bases in the two genomes differ: Fig. 6.1a shows a CG pair, while Fig. 6.1b shows a TG pair. When an SNP occurs, the complementary structures of the DNA mean that the other strand is always affected as well. Since the genome is diploid in the form of double chromosomes, SNPs occur as wild-type, heterozygous or homozygous.

There are approximately 15 million such genetic variants in the genome of an individual (Fig. 6.2). These variants may be unique to that individual or may be present in others. Some variants increase the risk of suffering from diseases, while others can reduce this risk. Others, however, do not affect at all on disease risk. In addition to single nucleotide polymorphisms, there is also deletion (missing DNA letter), insertion (additional insertion), and translocation (change of location/displacement) of DNA letters that influence genetic variation.

The question is how do these genetic variants influence the risk of certain diseases?

Today, genetic diseases are divided into two classes: Those associated with a single gene (monogenic: Mono = single; genic: Genetic), and those influenced by multiple genes

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_6

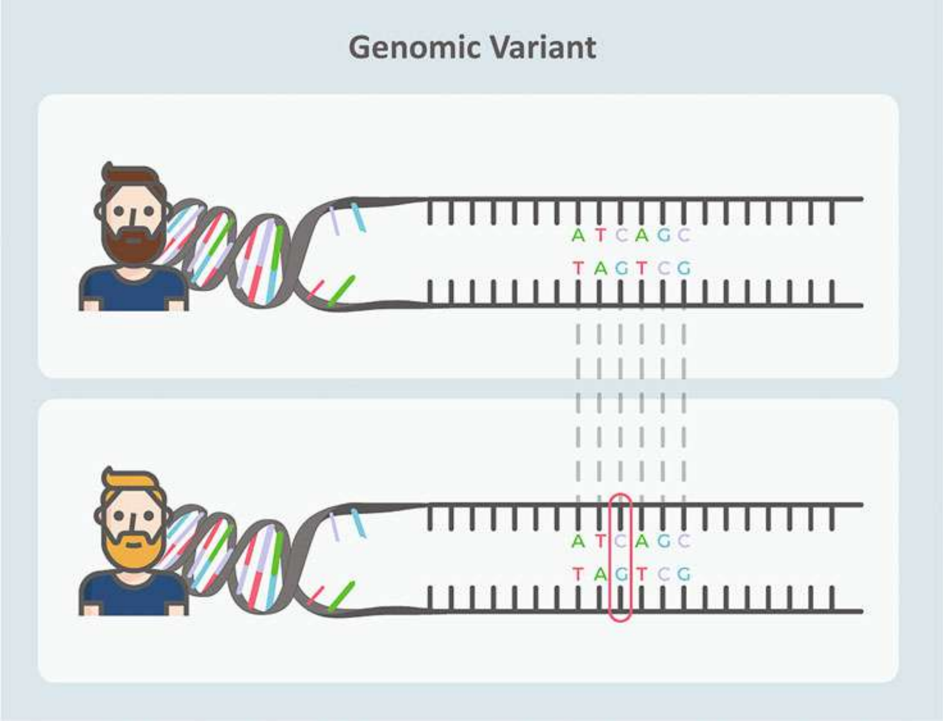


Fig. 6.1 Individual differences in the genome are called genetic variants. If they occur more frequently than 1%, they are referred to as single nucleotide polymorphisms. If they occur less frequently, they are referred to as mutations. (**a**) CG pair, (**b**) TG pair in the genome of two individuals

(polygenic; poly = many; genic = genetic) and environmental factors. Many diseases lie on a spectrum between these two extremes.

Many inherited diseases can be traced back to variants in a single gene (Fig. 6.3). For example, cystic fibrosis (or cystic fibrosis), a progressive genetic disease caused by a single mutation in a chloride ion channel (cystic fibrosis transmembrane conductance regulator—CFTR gene on chromosome 7), leads to lung infections and severely limits the ability to breathe, as the defective ion channel prevents the mucus from liquefying and keeps it viscous.

Diseases that are based on the defect of a gene are also called Mendelian diseases because they follow the Mendelian rules of inheritance.

Complex diseases, also called polygenic diseases, are caused by many genetic variants coupled with environmental influences (such as diet, sleep, stress, and smoking).

One example is coronary artery disease (CAD). To date, researchers have found about 60 genetic variants that are more common in people with coronary heart disease. Most of

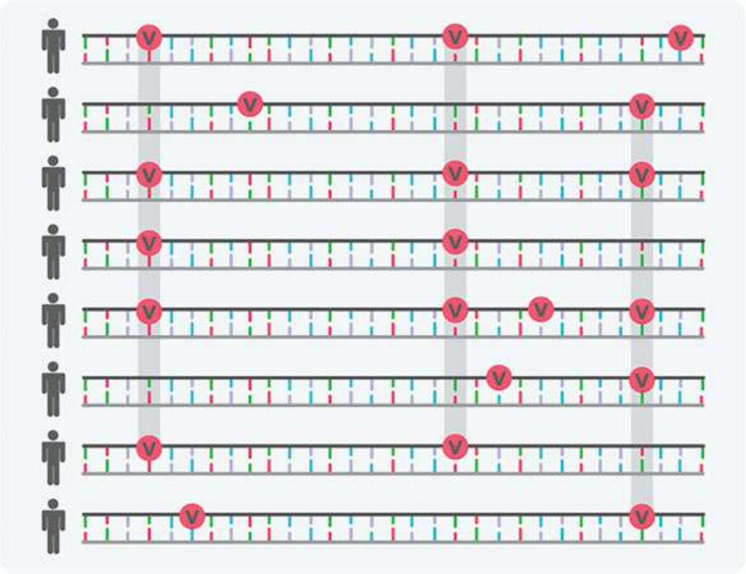


Fig. 6.2 Genetic variants are distributed throughout the genome

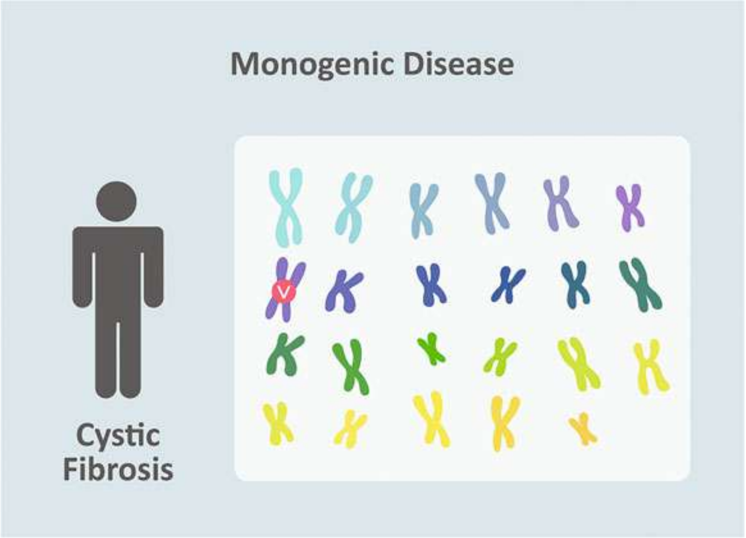


Fig. 6.3 In a monogenic disease, only one gene on one chromosome is affected

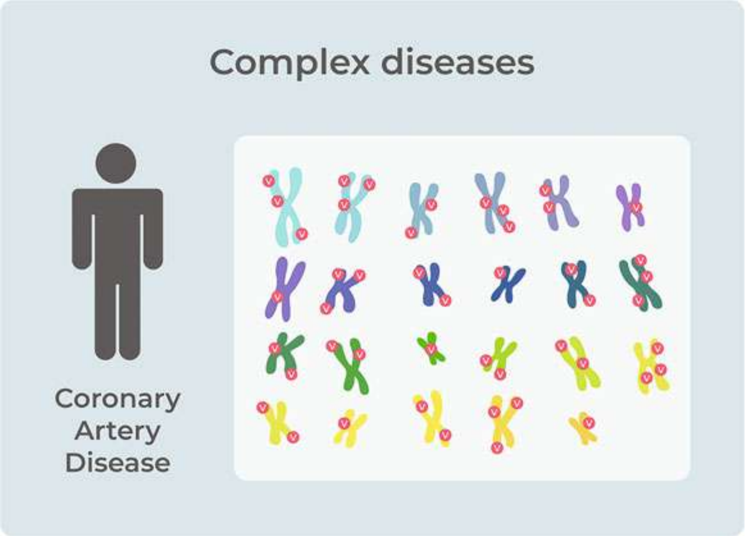


Fig. 6.4 A polygenic disease is often a complex interaction of many genes on different chromosomes

these variants are distributed across the genome and do not cluster on a particular chromosome (Fig. 6.4).

Researchers identify these genetic variants associated with complex diseases by comparing the genomes of individuals with and without these diseases, known as genome-wide association studies (GWAS).

When comparing genomes of healthy (controls) and diseased (cases) participants, the following terminology has evolved to describe genetic variants:

Reference Allele

The reference allele is the nucleotide that occurs in the reference genome used here. It is important to always name the specific reference genome referred to here. Example: Genome Reference Consortium GRCh38 from December 2013.

Variant Allele

The variant allele is a different nucleotide that occurs at that location in the genome compared to the reference allele. Usually, only one allele is observed that differs from the reference sequence. However, at some sites, several differences can be seen.

Effect Allele

The effect allele is the allele used to measure the size of the observed phenotypic effect.

(continued)

Minor Allele

The minor allele is the nucleotide that is least abundant at that position in the population under study. Usually, but not always, the minor allele is the same as the effect allele.

When comparing different populations, e.g. of European and Asian ancestry, the minor allele may differ. That is, an allele that is predominant in a European population may be a minor allele in an Asian population.

Furthermore, variant alleles show the following properties.

Allele Frequency

Each allele occurs in a population at a certain prevalence or frequency. Allele frequency is expressed as a fraction or percentage of chromosomes carrying the allele in the population under study. For example, an allele with a frequency of 0.2 is present in 20% of chromosomes. Since individuals can be either homozygous for the allele—carrying two copies—or heterozygous—carrying one copy—the proportion of individuals carrying an allele cannot be equated with allele frequency.

Minor Allele Frequency

Minor allele frequency (MAF) is the frequency of the minor allele in a given population. MAF is equivalent to effect allele frequency when the minor and effect alleles are the same. Variants with MAF greater than 5% are considered frequent; variants with MAF between 0.05% and 5% are called low frequency. Variants with a MAF of $< 0.05\%$ are designated as rare.

Linkage Disequilibrium

Linkage disequilibrium (LD) means that two alleles at different gene loci occur together less frequently or more frequently than would be the case by pure chance.

According to Mendel's rules, genes are passed on independently to the next generation. However, this may not be the case if the genes in question are located on the same chromosome or close to each other on the same chromosome. Then, the probability that they will be inherited together is much higher. However, this is not always the case. Since the chromosomes attach to each other during meiosis, it can happen that individual chromosome arms overlap, break apart, and then become incorrectly linked again. This is called crossing-over. In this process, genes that were previously located on one chromosome can be separated. Therefore, the closer two genes are to each other, the more likely they are to be inherited together and the less likely they are to be separated by crossing-over.

LD is usually defined as r^2, which examines the allele frequency of the two alleles under consideration. An r^2 value of 1 indicates that the alleles are completely correlated. They are always inherited together. An r^2 value of 0 says that the alleles are completely independent of each other. They are inherited independently of each other. Usually, linkage disequilibrium can be quickly estimated from the minor allele frequency (MAF) of the gene variants under study. If the value of the MAF is approximately the same for both gene variants, they are inherited together.

6.1 Genome-Wide Association Studies

A genome-wide association study (GWA study or GWAS) is an observational study of a genome-wide set of genetic variants in different individuals. The aim is to determine whether one or more gene variants are associated with a trait or disease. GWAS usually focuses on associations between single nucleotide polymorphisms (SNPs) and traits such as severe human diseases but can equally be applied to other genetic variants and other organisms.

A genome-wide association study is consistent with the forward genetics approach.

Specifically, for cost reasons, one to two million gene variants that have been characterized by a genotyping chip (SNP microarray) for the individual participant will be examined.

If there is a statistically significant difference (excess probability p for Latin probilitas = probability: p-value below a certain threshold) between the frequency of an allele in cases compared to controls, the gene variant is associated with the disease. For visualization purposes, the p-values for the individual gene variants are plotted on a graph in the order of their location on the chromosomes in a so-called Manhattan plot (named after the skyline of the New York borough of Manhattan) (Fig. 6.5).

The most widely used program for analyzing a genome-wide association study is PLINK (https://www.cog-genomics.org/plink/2.0/).

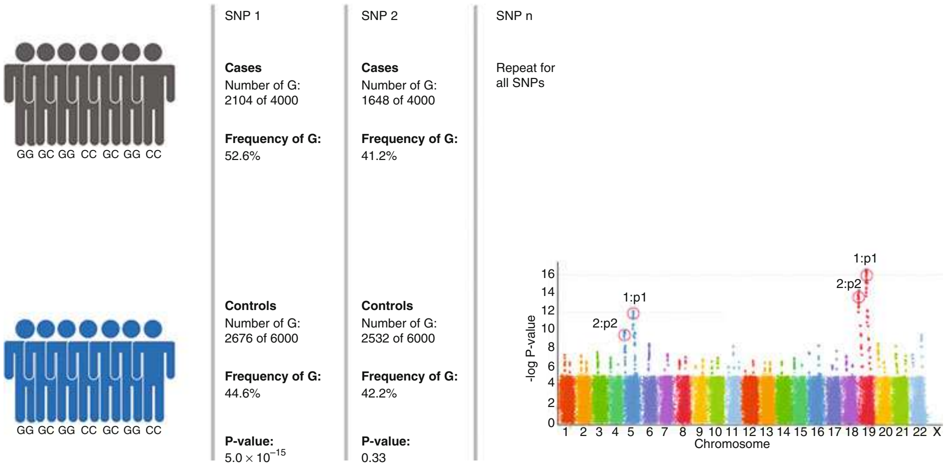


Fig. 6.5 Illustration of the principle of the genome-wide association study (GWAS). A population of cases is contrasted with a population of controls. For each gene variant, the frequency in the respective population is calculated and compared. The exact calculation of the individual p-values can be found in the example

Example (by Alicia Martin, Stanford University)

Calculation of the p-value for the gene variant rs456789 using 12 diseased (“cases”) and 24 healthy controls (“controls”):

Status	AA	AG	GG
Healthy	2	3	7
Sick	4	9	11

1. Counting alleles: There are 2 As for AA homozygous carriers and 1 A for heterozygous carriers. The same is true for G for AG heterozygotes (1) and GG homozygotes (GG).

	Sick	Healthy	Total
A	$2 \cdot 4 + 1 \cdot 9 = 17$	$2 \cdot 2 + 1 \cdot 3 = 7$	24
G	$2 \cdot 11 + 1 \cdot 9 = 31$	$2 \cdot 7 + 1 \cdot 3 = 17$	48
Total	48	24	72

2. Normalization of the sums gives the allele frequencies (a division of the number of alleles by the total number of alleles):

	Sick	Healthy	Total
A	17/72 = 23.6%	7/72 = 9.7%	33.3%
G	31/72 = 43.1%	17/72 = 23.6%	66.7%
Total	66.7%	33.3%	100%

3. Calculation of the expected allele frequencies:

	Sick	Healthy	Total
A	$0.33 \cdot 0.667 = 22.2\%$	$0.333 \cdot 0.333 = 11.1\%$	33.3%
G	$0.667 \cdot 0.667 = 44.5\%$	$0.333 \cdot 0.677 = 22.2\%$	66.7%
Total	66.7%	33.3%	100%

4. Calculation of the expected number of alleles (multiplication of the frequency by the total number of alleles):

	Sick	Healthy	Total
A	$0.222 \cdot 72 = 15.98$	$0.111 \cdot 72 = 8.00$	23.98
G	$0.445 \cdot 72 = 32.04$	$0.222 \cdot 72 = 15.98$	48.02
Total	48.02	23.98	72

5. The chi-square test (χ^2-test) (Fig. 6.6) is used to check in which way the alleles are distributed in order to assess how statistically valid the statement is:

B = observed, E = expected $\chi^2 = \sum \frac{(B-E)^2}{E.}$

$$\chi^2 = \frac{(17 - 15.98)^2}{15.98} + \frac{(7 - 8)^2}{8} + \frac{(31 - 32.04)^2}{32.04} + \frac{(17 - 15.98)^2}{15.98}$$

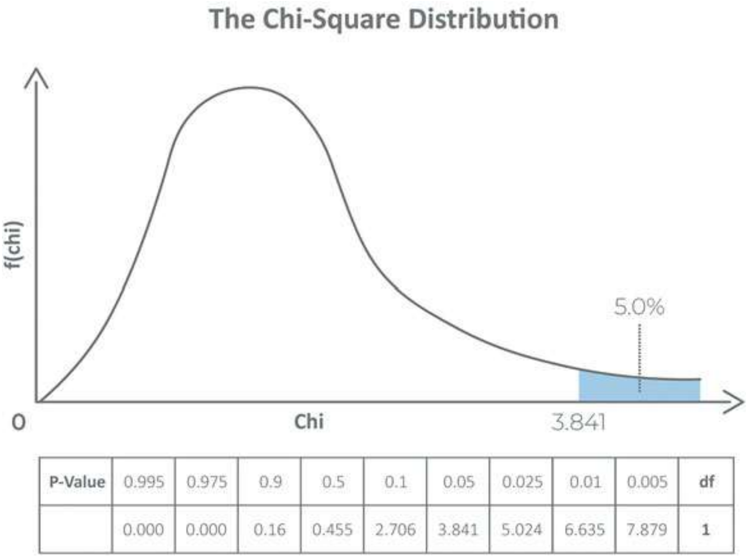
$= 0.07 + 0.13 + 0.03 + 0.07$
$= 0.3$

In the next step, the p-value can be estimated using the chi-square distribution. The number of degrees of freedom (df) is calculated according to the number of classes (k) minus 1:

$\mathrm{df} = \mathrm{k} - 1$

In this case, two classes exist (sick and healthy), which corresponds to a degree of freedom of 1 according to the above formula. In the chi-square distribution (Fig.7.2), a chi-square value of 0.3 then corresponds to a p-value of 0.58:

$p(\chi^2 = 0.3) = 0.58$



P-Value	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
	0.000	0.000	0.16	0.455	2.706	3.841	5.024	6.635	7.879	1

Fig. 6.6 Chi distribution for determining the p-value

The P-Value

The p-value, or significance value, is a measure of credibility. The smaller the p-value, the less likely it is that the gene variant is NOT associated with the disease. In general, the rule is that the p-value should be less than the significance level alpha of 0.05. In other words, 95% correct. However, in a GWA study, you test an average of one million SNPs with far fewer cases and controls.

Alpha Error Accumulation in High-Dimensional Data

With high-dimensional data (a small sample of thousands of patients in comparison with millions of gene variants), there is a risk of the alpha error accumulation problem (Multiple Testing Problem): The more hypotheses (SNPs) one tests on a data set with few patients, the higher the probability that one of them will be tested as false positive. In our case of one million SNPs, there is a high probability that several SNPs show an association with the disease only by chance in a smaller sample. To exclude this, we use the Bonferroni correction: If we test n independent hypotheses (SNPs) on a data set, the statistical significance is 1/n times the significance that would result from testing only one hypothesis. In our case, where 500,000 SNPs have been tested, one needs a p-value of 0.05/500,000 = $10^{-7.}$ Returning to our example above, it shows that rs456789 is not significantly associated with the disease.

Odds Ratio

After finding a SNP that is significantly related to the disease, we can calculate the odds ratio or chance ratio.

The probability P (Probability) of being ill when carrying allele A is:

$$\mathrm{P(sick|\ A)}$$

Where the probability P of being sick and carrying allele A is divided by the probability of carrying allele A:

$$\mathrm{P(sick|A)} = \frac{\mathrm{P(sick\&A)}}{\mathrm{P(A)}}$$

Next, we can convert this probability into odds or odds by dividing it by (1 minus itself) (75% chance means a 3:1 chance of it happening):

$$\mathrm{Odds(A)} = \frac{\mathrm{P(sick|A)}}{1 - \mathrm{P(sick|A)}}$$

Then, we infer the odds ratio by dividing the odds of both genotypes:

$$\text{Odds Ratio} = \frac{\text{Odds (A)}}{\text{Odds (B)}}$$

$$\text{P(sick|A)} = \frac{17}{(17+7)} = 0.708$$

$$\text{P(sick|G)} = \frac{31}{31+7} = 0.816$$

$$\text{Odds(A)} = 2.42$$

$$\text{Odds(G)} = 4.43$$

$$\text{Odds Ratio} = \frac{\text{Odds (A)}}{\text{Odds (G)}} = 0.55$$

The odds ratio (OR) is used to represent the magnitude of the association between an allele and a binary disease (diseased or healthy) or trait. An odds ratio of:

- **Close to 1** means that the allele has little or no effect on the disease.
- **Greater than 1** means that the effect allele is more likely to be carried by people with the condition.
- **Less than 1** indicates that the effect allele is more likely to be carried by people without the disease, suggesting that the allele is protective.

For our example above, an odds ratio of 0.55 does not specifically mean that they are 0.55 times less likely to have the disease. It means that the A allele is not significant for the disease.

The Effect Size (Beta)

The effect size (beta) is analogous to the odds ratio but can be used on continuous characteristics such as disease severity but better for characteristics, such as weight or cholesterol level. It represents the magnitude and direction of the association between a variant and that trait.

The Confidence Interval (CI)

The confidence interval (CI) is the range of expectation for the odds ratio and effect size that includes the true location of the parameter with a certain probability (the confidence level) when the experiment is repeated infinitely.

Example: 95% CI for an odds ratio of 0.852 to 0.941 means that there is a 95% chance that the odds ratio is between 0.852 and 0.941.

6.2 Fine Mapping

While the whole genome sequencing from the participant's blood or saliva sample to the genome sequence file still costs on average more than 1000 euro, the cost from sample to data is less than 25 euro for a SNP microarray. Consequently, large populations are only genotyped instead of sequenced, despite advances in whole-genome sequencing.

In designing genotyping arrays developed for GWAS, one cleverly uses linkage disequilibrium (LD) to cover the entire genome by genotyping a subset of variants. Therefore, it is important to know: It is very unlikely that the associated gene variants found actually correspond to the causal gene variants. The regions on the chromosome that are in LD and therefore inherited together are also called haplotype blocks (Fig. 6.7). These blocks often contain hundreds of gene variants encompassing many different genes. The latter poses a problem in that it complicates the biological interpretation of a gene variant found since it cannot be assigned to one gene.

For this reason, the haplotype block with the gene variant found is sequenced in the entire population studied. This is referred to as fine mapping in order to be able to identify the causal gene variant beyond the associating gene variant. Ultimately, this approach is still significantly cheaper than performing whole-genome sequencing (WGS) in large populations.

To query whether one or more SNPs are in linkage disequilibrium (LD), the following website can be used to quickly query a population: https://ldlink.nci.nih.gov/?tab=home.

6.3 Polygenic Risk Score

A polygenic risk score (PRS) is a number based on the variation of multiple genetic loci and the associated probabilities from regression analysis (see the example in 6.1). It serves as the best prediction for the disease or trait that can be made given the variation in multiple genetic variants.

The background to the polygenic risk score is the fact that genetically complex diseases or traits cannot be attributed to one or a small number of gene variants. For this purpose, the sum of all odds ratios (in the case of binary diseases or traits) or effect sizes (in the case of non-binary diseases or traits) of the gene variants that were identified as significant in a regression analysis is formed.

$$\mathrm{PRS}_j = \sum^{n} \left[\beta_{i,\mathrm{discovery}} \cdot \mathrm{SNP}_{ij}\right]$$

The polygenic risk score is the sum of all gene variants for which

- $\beta_{i,\mathrm{discovery}}$ = effect size (continuous characteristic) or an odds ratio (binary characteristic;$\beta = \log(\text{odds ratio})$ is present).

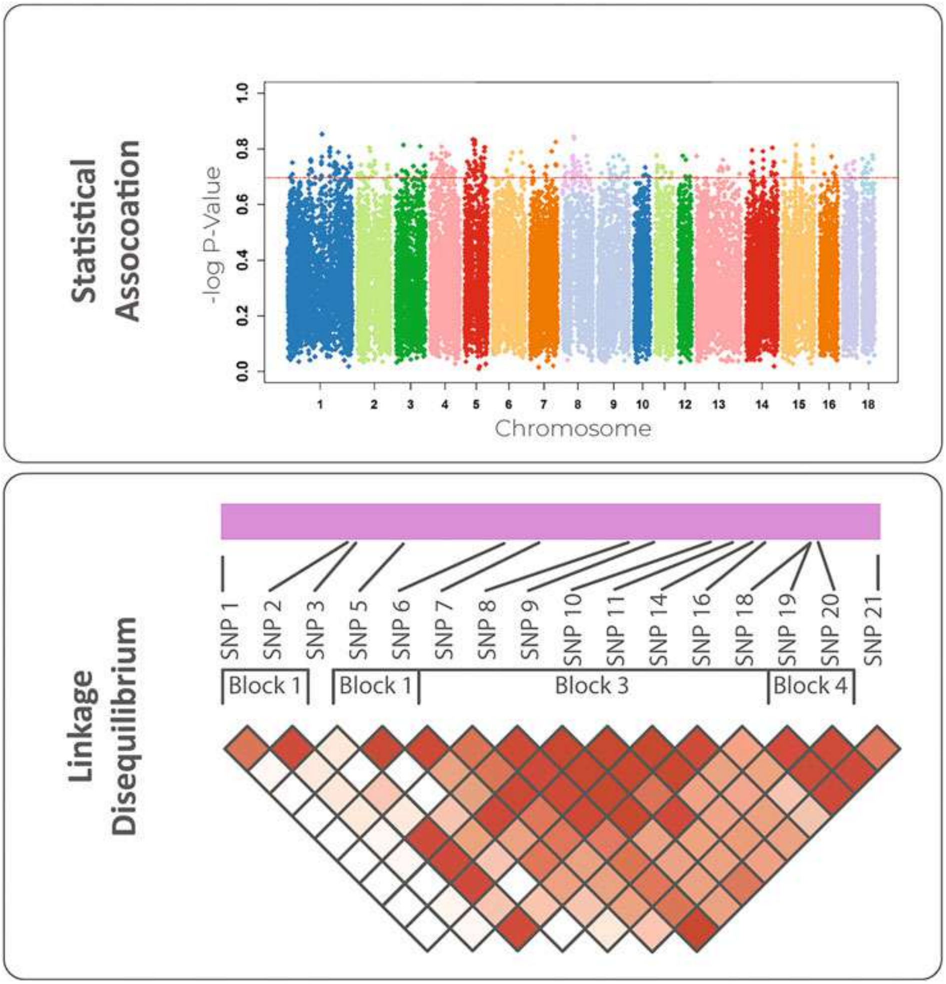


Fig. 6.7 Gene variants are not inherited independently of each other. Rather, the gene variants are located in so-called haploblocks on the chromosome. The closer two gene variants are located to each other, the more likely they are to be inherited together

- SNP_{ij} = number of alleles (0, 1, 2) for SNP *i* of individuals *j* in the population under study.
- n = number of SNPs

Polygenic risk scores are, therefore, purely **additive** prediction models. The PRS does not address gene-gene interactions, also known as epistasis, i.e. the fact that a disease or a trait only develops in combination with certain gene variants.

Polygenic risk scores (PRS) are primarily used in plant breeding. With the large sequencing projects in human populations, PRS is now being used for risk detection in humans. The problem with PRS is that they are usually formed on several hundreds of thousands of gene variants. PRS, formed on GWA studies in the US, do not apply to Asian or European populations. Due to a large number of gene variants, there is often overfitting of the PRS for the underlying population. Overfitting refers to a specification of the model for the test population that contains too many explanatory variables (gene variants) and is therefore not applicable to other populations. A further disadvantage associated with overfitting in the form of the sum of hundreds of thousands of gene variants is that the one gene whose product is suitable as a drug target is not identified for drug development (see 6.4).

To create a polygenic risk score (PRS), one can access collected p-values at GWAS Catalog: https://www.ebi.ac.uk/gwas/ and use programs such as LDPred: https://github.com/bvilhjal/ldpred to calculate the score.

6.4 Drug Target Linkage

While GWAS studies have succeeded in linking—and more importantly, decoupling—genes and disease, much of the opportunity in human genetic analysis arises not from the study of "common" variations but "rare" variants found in only a few people. These rare variants are of interest for two reasons: First, common variants tend to have little impact on disease pathology compared to rare variants, which often have a much greater clinical impact. Rare variants are therefore more informative in terms of disease and can be very informative regarding the molecular roots of the disease. They offer deeper insights into molecular pathways and their foci, potential drug targets, and opportunities to identify novel molecular biomarkers and disease subtypes. Second, beyond individual variations, it has been recognized that it is possible to infer the efficacy of a drug candidate from these. Genetic links to disease predict the efficacy potential of the drug target.

Several examples support this approach, including sclerostin, SCN9A, ANGPTL4, and, most importantly, PCSK9. Interest in PCSK9 as a drug target began with the discovery of rare gene variants that either increased LDL cholesterol and heart disease risk or decreased both. Importantly, a person who carried two completely deactivated copies of the PCSK9 gene (equivalent to complete inhibition of the PCSK9 protein by a drug) had no other related health problems. This indicates that the benefits of low LDL cholesterol did not cause other health problems. Thus, a "clean" inhibitor of PCSK9 could be both effective and safe, as shown later (Figs. 6.8 and 6.9).

What does this mean for drug discovery? In recent years, several large retrospective analyses, as Table 6.1 shows, have investigated whether drug candidates bind to drug targets that are genetically linked to the disease (drug target linkage) (Fig. 6.8), showing a higher success rate than drugs that bind to drug targets that do not show a genetic link to the disease. Thus, the effect of drug target linkage was shown to be greatest, particularly in phase II clinical trials. In this phase, the clinical effect is measured for the first time. Using the public financial records of listed pharmaceutical companies, it was even possible to calculate cost savings of up to US$24.6 million per drug candidate and per year.

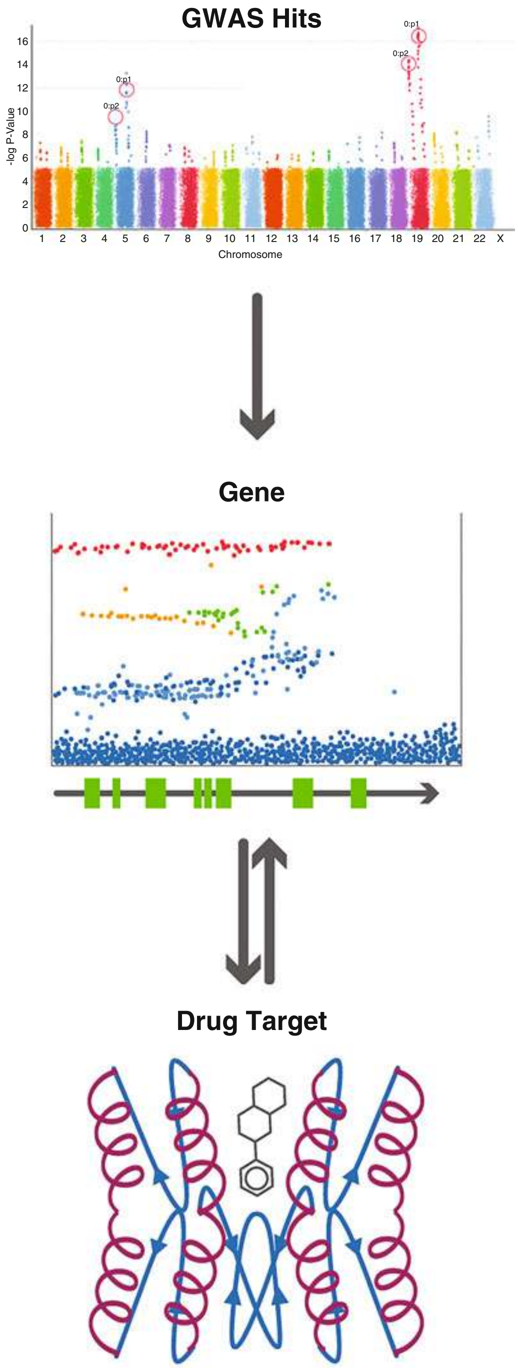


Fig. 6.8 Illustration of the genetic linkage of a drug target to a disease. Genes characteristic of the disease are found in a genome-wide association study (GWAS) (see Manhattan plot above). Structural analysis of chromosome localization suggests a specific gene whose gene product is a drug target

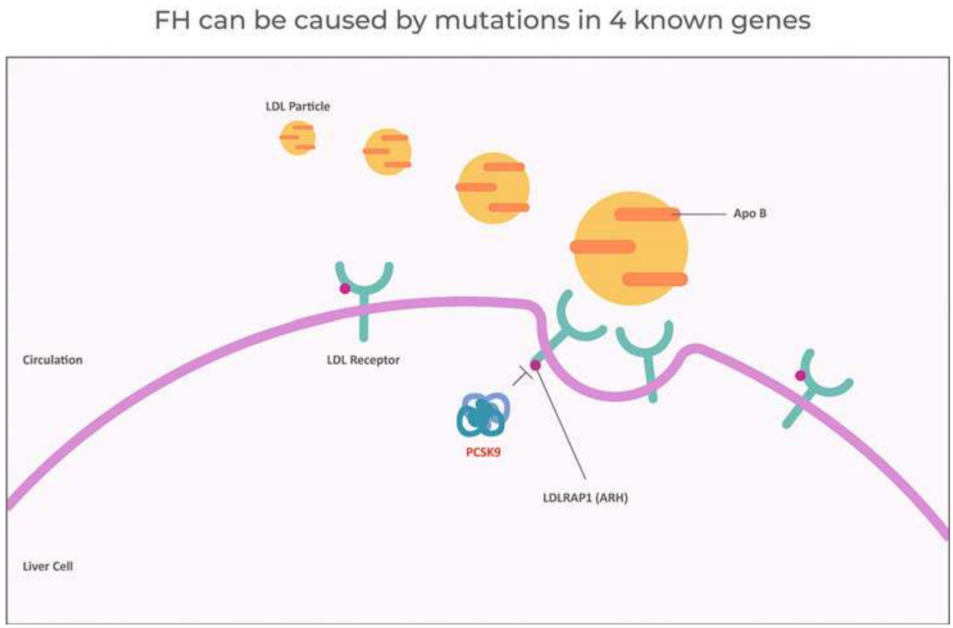


Fig. 6.9 Gene variants in PCSK9, which prevent the function of the gene product, a serine protease, were found in patients who had particularly low blood LDL cholesterol. Closer biochemical studies revealed that the serine protease PCSK9 minimizes the number of LDL receptors, thereby preventing the uptake of LDL cholesterol bound by apoB. Inhibiting PCSK9, the number of LDL receptors on the surface of liver cells remains constant, more LDL cholesterol is taken up from the blood. PCSK9 is the best-known example where a drug target was discovered in a GWA study. Currently, two drugs that inhibit PCSK9 are approved for familial hypercholesterolemia (FH)

Table 6.1 Success rate of drug development in the respective phase or overall

	Preclinical	Phase I	Phase II	Phase III	Expert opinion
Success rate	65%	45%	25%	65%	85%
Success rate if the genetic link to drug target exists	k. A.	52–62%	38–49%	67–80%	k. A.
Cost savings US$ per chemical probe and year	k. A.	1.1–2.75	13.8–24.6	0–18.9	k. A.

6.5 Phenomenon-Wide Association Study—PheWAS

In the first part of this book, we had already learned about the forward genetics approach. Genome-wide association studies (GWAS) follow this strategy. However, instead of studying one gene, one studies genomes, as the interaction of all genes, for phenotypes

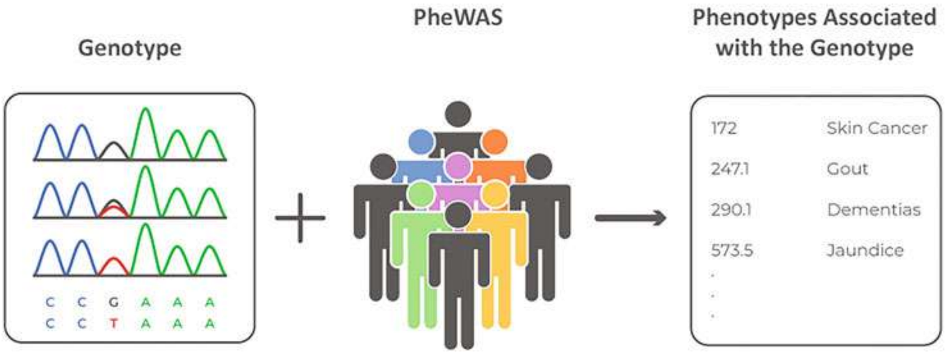


Fig. 6.10 Illustration of the principle of a phenome-wide association study (PheWAS): Instead of looking at just one phenotype, the association in a large number of phenotypes is considered for each gene variant. The aim is to find gene variants or genes that are particularly central for several phenotypes such as diseases or traits

such as a disease or trait. We also learned that GWA studies involve several thousand participants but examine significantly more gene variants (500,000 to one million SNPs). In technical jargon, such a high-dimensional data structure is called wide data, in distinction from big data. If a large number of gene variants are tested in a population that is significantly smaller than the number of gene variants to be tested, associations are found by chance. This is also known as the multiple testing problem. In GWA studies, the Bonferroni correction for associating gene variants is therefore used to ensure that the association is strong enough not to have been found by chance.

Another exciting approach to tackle this problematic data structure is to follow the backward genetics approach. This is called the Phenome-wide association study (PheWAS), in contrast to genome-wide association studies. Basically, it is assumed that gene variants that are essential for the occurrence of a disease or trait must also do so in other diseases or traits. Indeed, as shown in Fig. 6.10, some gene variants show associations across multiple diseases and populations. The genes and especially their products, the proteins, of these gene variants are readily used in drug discovery as a starting point for drug development. From an economic point of view, the risk is lower if the chemical probe for the addressed chemical probe target later allows several indication areas.

Summary

Genomic variants can be the cause of diseases. Some variants show a large effect, as they are all responsible for a disease. In contrast to these monogenic diseases, there are a large number of complex diseases that are based on an interaction of many genes and the environment. To study complex diseases, genome-wide association studies (GWAS) are performed. Here, each gene variant is examined for its effect on the observed phenotype.

Multiple gene variants are combined into a polygenic risk score (PRS) to predict polygenic diseases or traits. Special attention for drug development is given to gene variants that can be assigned to the disease mechanism and whose gene products can serve as drug targets. The best-known example is PCSK9 for lowering excessive LDL cholesterol.

Further Reading

Abifadel M et al (2003) Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nat Genet 34:154–156

Bush WS, Oetjens MT, Crawford DC (2016) Unravelling the human genome-phenome relationship using phenome-wide association studies. Nat Rev Genet 17:129–145

King EA, Davis JW, Degner JF (2019) Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. PLoS Genet 15:e1008489

Nelson MR et al (2015) The support of human genetics evidence for approved drug indications. Nat Genet 47:856–860

Purcell S et al (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81:559–575

Vilhjalmsson BJ et al (2015) Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am J Hum Genet 97:576–592

Visscher PM et al (2017) 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet 101:5–22

Gene Therapy and Genome Editing 7

In recent years, scientific successes have increasingly brought to the fore the possibilities of modifying genes and the genome as a therapeutic intervention. Specifically, gene therapy drugs such as Zolgensma have been approved for the treatment of hereditary spinal muscular atrophy, and new genome editing techniques such as CRISPR-Cas have been developed. So, what is the difference between gene therapy and genome editing?

Gene Therapy
Gene therapy is a process in which one or more copies of a gene are introduced into a patient's body using a vehicle. This approach repairs the effect of the faulty gene causing a genetic disease or condition. To have a lasting effect, the introduced genes must be expressed over an extended period (ideally the patient's entire life).

Genome Editing
In contrast, genome editing aims to permanently correct or remove disease-causing genes. Such correction remains throughout the life of the transformed cells. Depending on the type of cells involved in the treatment and the lifespan of those cells in the body, genome editing treatments can effectively treat disease with either a single application or multiple repeated applications.

7.1 Gene Therapy

Gene therapy refers to the successful transfer of a gene into the patient's cell so that this gene is read there (Fig. 7.1). The goal is for patients with a defective, disease-causing gene to express an intact gene (Fig. 8.1). Introducing an intact gene into the patient's cell is not easy. Usually, retroviruses or adenoviruses are genetically modified and used as

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_7

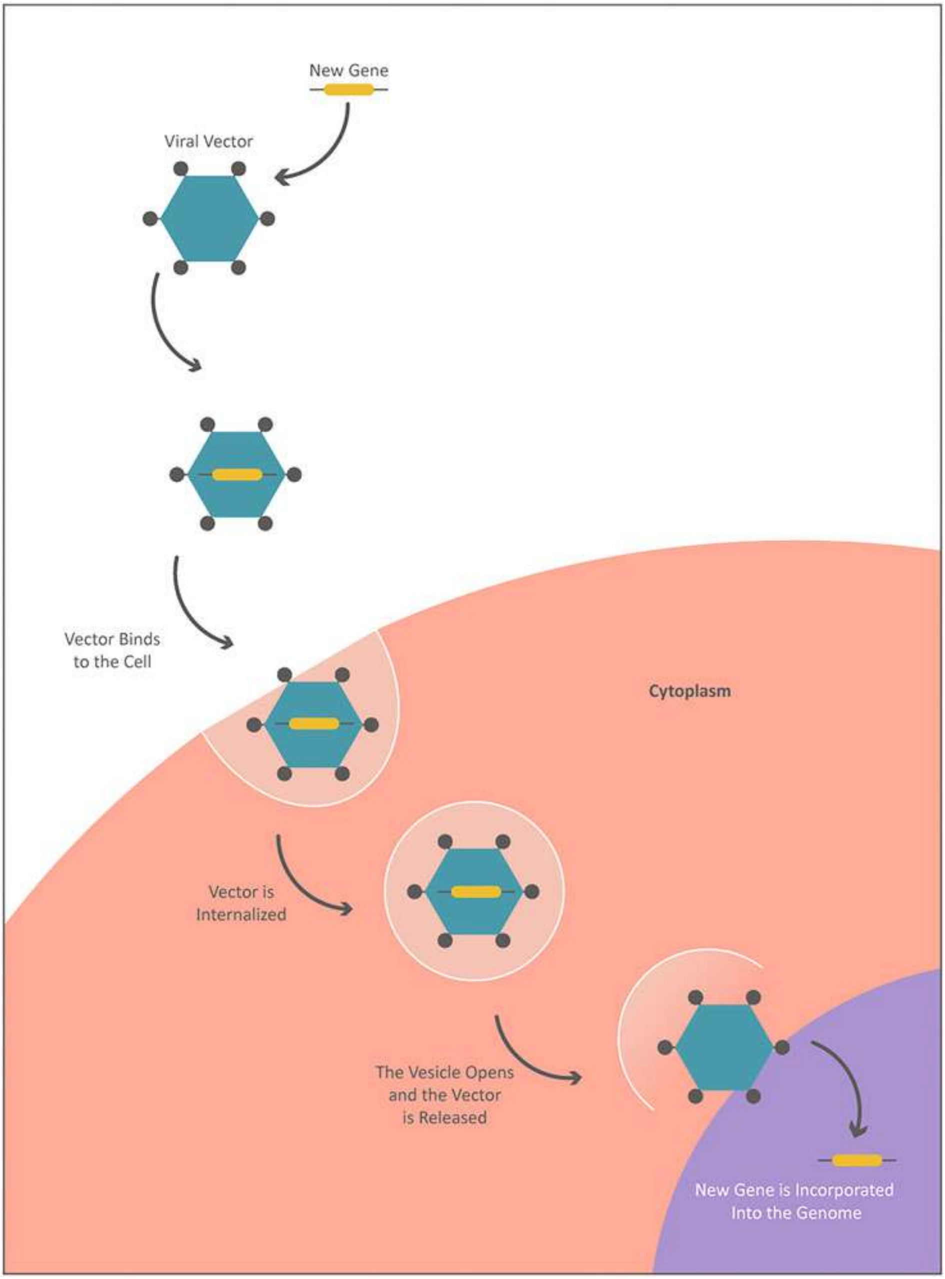

New Gene
Viral Vector
Vector Binds to the Cell
Cytoplasm
Vector is Internalized
The Vesicle Opens and the Vector is Released
New Gene is Incorporated Into the Genome

Fig. 7.1 Illustration of the principle of gene therapy based on the introduction of a gene into the cell

transporters of the intact gene. Currently, the best-known gene therapy is Zolgensma for the treatment of spinal muscular atrophy (SMA). SMA is a neuromuscular disorder caused by a mutation in the SMN1 gene that leads to a decrease in the SMN protein, which is required for motor neuron survival. Zolgensma is, at its core, an intact SMN1 gene encased in an adeno-associated virus (AAV) capsid. The adeno-associated virus acts as a carrier and is also known as a vector. The AAV vector delivers the intact SMN1 gene to the affected motor neurons, where it leads to an increase in SMN protein.

The procedure generally has disadvantages: On the one hand, the intact gene is not incorporated into the host genome. Thus, the gene is only temporarily present in the cell and further gene therapy is necessary later. On the other hand, the transferred gene is incorporated into the host genome. However, it is not known where and what consequences this might have. In addition, gene therapy is currently limited to monogenic diseases. The transfer of a gene and the correction of a defective gene are already a challenge. As the number of genes involved increases, the complexity increases, which has so far prevented the development of gene therapies of complex genetic diseases.

Gene therapy can be divided into two types:

Somatic Gene Therapy

In gene therapy with somatic cells, the therapeutic genes are transferred into a cell other than the germ cell or an undifferentiated stem cell. Such changes affect only the individual patient and are not passed on to offspring. Somatic gene therapy is the currently approved treatment for diseases.

Germline Gene Therapy

In germline gene therapy (GGT), germ cells (sperm or eggs) are modified by introducing functional genes into their genomes. The modification of a germ cell causes all cells of the organism to contain the modified gene. The modification is therefore heritable and is passed on to later generations. Many countries, including Germany, have banned germline gene therapy in humans for ethical reasons.

Somatic gene therapy is divided into in vivo and ex vivo. In the in vivo procedure, the gene transfer vehicle is injected directly into the patient, whereas in the ex vivo procedure, cells are removed from the patient and the gene transfer takes place in cell culture in the laboratory. Only after successful expression of the gene are the cells reimplanted into the patient (Fig. 7.2). All currently approved gene therapies are ex vivo procedures, as gene transfer into germline cells can be excluded here.

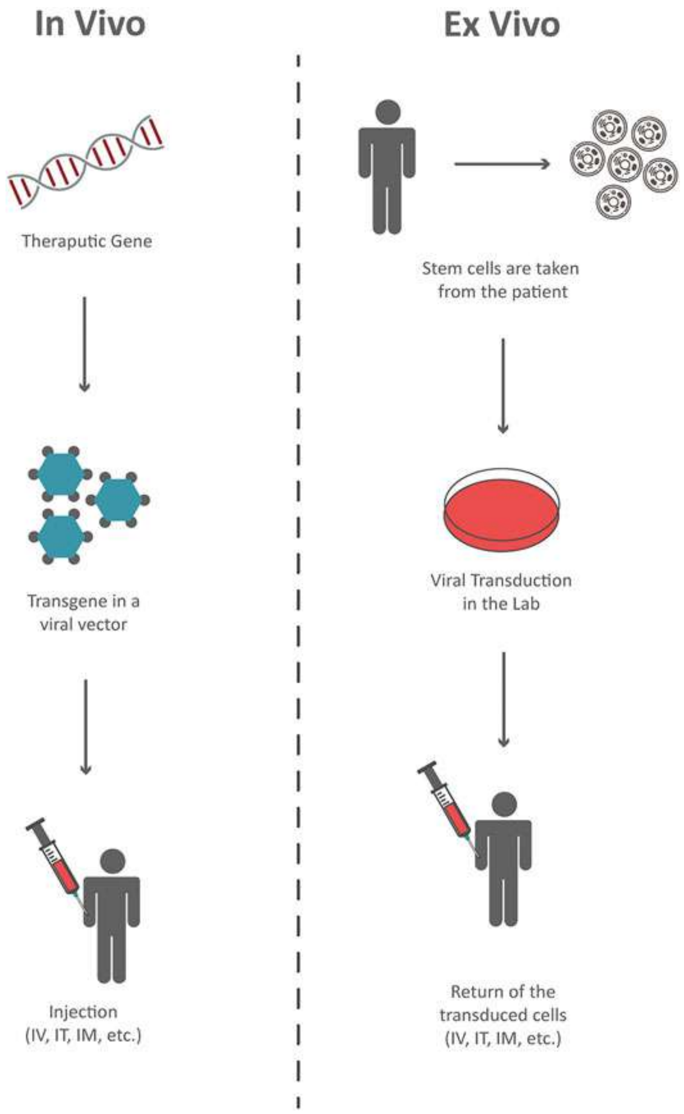

In Vivo
Ex Vivo
Theraputic Gene
Transgene in a viral vector
Injection (IV, IT, IM, etc.)
Stem cells are taken from the patient
Viral Transduction in the Lab
Return of the transduced cells (IV, IT, IM, etc.)

Fig. 7.2 Difference between in vivo and ex vivo gene therapy

7.2 Genome Editing

Genome editing is a genetic engineering technique in which DNA is inserted, deleted, modified, or replaced in the genome of a living organism. Unlike other genetic engineering techniques in which genetic material is randomly inserted into a host genome, genome editing aims to modify (edit) at site-specific locations.

The best-known genome editing method is CRISPR-Cas. Originally, clustered regularly interspaced short palindromic repeats (CRISPR) are genetic elements that bacteria use as a kind of acquired immunity to protect themselves from viruses. They consist of short sequences derived from viral genomes that have been incorporated into the bacterial genome. These short sequences are transcribed. The nuclease CRISPR-associated protein (Cas) can bind these sequences. Subsequently, the sequences serve Cas as a template to recognize and cut complementary viral DNA sequences. CRISPR can be thought of as the acquired immune system of a bacterium.

In genome editing, the nuclease has been genetically modified so that it binds a synthetic guide RNA (gRNA) and, mediated by the gRNA, specifically cuts at a specific site in the genome as shown in Fig. 7.3. In the process, existing genes can be removed or new ones can be inserted at the site.

The specificity of editing depends on two factors: The target sequence of the guide RNA (gRNA) and the PAM sequence. The gRNA is 20 bases long. The Cas nuclease binds to the correct location in the host genome using the gRNA to bind to base pairs on the host DNA. The sequence is not part of the Cas protein and is therefore customizable and can be synthesized independently. The Cas nuclease is nevertheless not completely free to bind anywhere in the genome based on the gRNA. The host genome must carry a short PAM sequence. The Cas nuclease cannot be easily modified to recognize a different PAM sequence. However, the PAM sequence is ultimately not limiting because it is typically a very short and nonspecific sequence that frequently occurs at many sites throughout the genome. The 5′-NGG-3′ PAM sequence occurs approximately every eighth to 12th base pair in the human genome.

The Cas9 nuclease works like genetic scissors, opening both strands of the target sequence of DNA to introduce the modification by one of two methods. As shown in Fig. 7.3, once the gene is excised, deletion occurs through DNA repair mechanisms that result in the two ends of the cut DNA being reconnected. Alternatively, the insertion of a new DNA strand occurs analogously via existing DNA repair mechanisms. This is called homology-directed repair because the DNA to be inserted has homologous overlaps on the left and right with the cut DNA ends of the host genome.

Summary

The modification of the genetic material for therapeutic purposes is carried out either with gene therapy or with genome editing. In gene therapy, a copy of a gene is introduced into the patient’s cells using a vehicle. The goal of gene therapy is to replace a faulty gene. Several gene therapies are approved today. In contrast, genome editing involves the addition, deletion, modification, or replacement of the genome. Genome editing has gained prominence, especially with the discovery of the CRISPR-Cas procedure. Genome editing

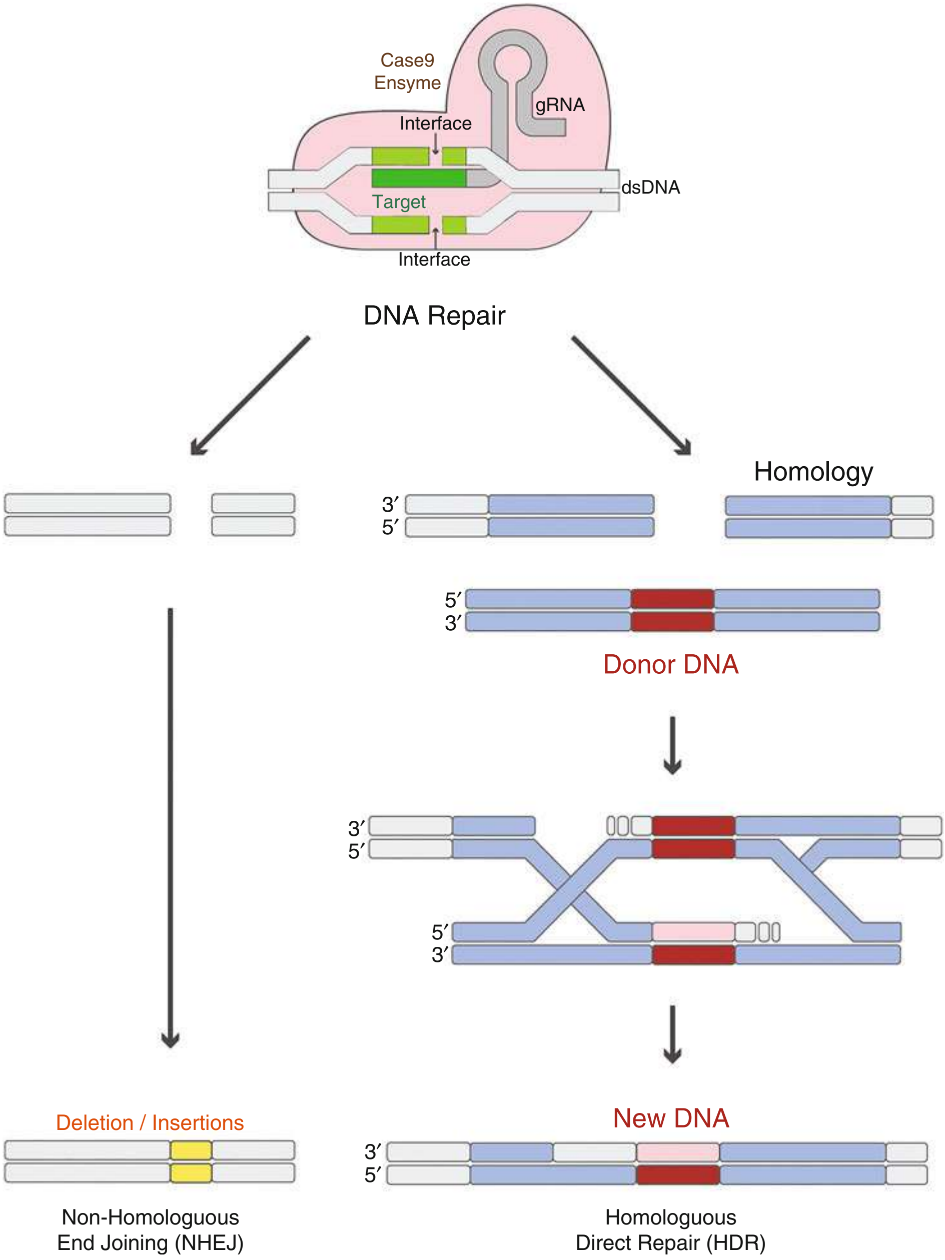


Fig. 7.3 Illustration of genome editing with CRISPR-Cas. The Cas9 nuclease (top) works like genetic scissors, opening both strands of the target sequence of DNA to introduce the modification by one of two methods. Deletion of the gene occurs after it is cut out by DNA repair mechanisms that result in the two ends of the cut DNA being reconnected. Alternatively, the insertion of a new DNA strand occurs analogously via existing DNA repair mechanisms. This is called homology-controlled repair since the DNA to be inserted has homologous overlaps on the left and right with the cut DNA ends of the host genome

is currently still an experimental procedure. An approved therapy based on this procedure does not yet exist.

Further Reading

Gaj T, Gersbach CA, Barbas CF (2013) ZEN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol 31:397–405

Jinek M et al (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337:816–821

Sheridan C (2011) Gene therapy finds its niche. Nat Biotechnol 29:121–128

Part III

RNA

RNA: Information and Function Carrier

8

In the molecular biology dogma, ribonucleic acid (RNA) is the mediator between the information carrier DNA and the functional carrier protein (Fig. 8.1). RNA is therefore often referred to as a hermaphrodite in the dogma since RNA is both an information carrier and a functional carrier. The information is carried by the complementary sequence of DNA. The function results from the fact that the ribosome, which converts messenger RNA into proteins, consists mainly of RNA.

RNA plays (almost) no role at all in drug development. Some natural substances are known that bind specifically only to bacterial ribosomes and inhibit them, and are therefore used as antibiotics. There have been considerations for some time to address mRNA as a drug target in order to inhibit translation and thus the formation of the target protein. Messenger RNA molecules are huge entities to which classical small-molecule drugs cannot specifically bind, let alone specifically inhibit. Therefore, oligonucleotides that can bind to their target mRNA in a sequence-specific manner have been developed in the past. Nevertheless, the oligonucleotides are usually too short to block the entire mRNA. Nevertheless, there are successful mRNA inhibitors such as mipomersen, which have been approved as drugs. For example, mipomersen initiates ribonuclease H by binding to mRNA, which leads to degradation of mRNA. However, the use of ribonuclease H-coupled mRNA inhibitors is limited in a sequence-specific manner.

The discovery of RNA interference (RNAi), a mechanism of gene regulation based on non-coding RNAs, electrified the pharmaceutical industry from the turn of the millennium. Cell experiments showed that small double-stranded oligonucleotides could specifically silence entire genes.

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_8

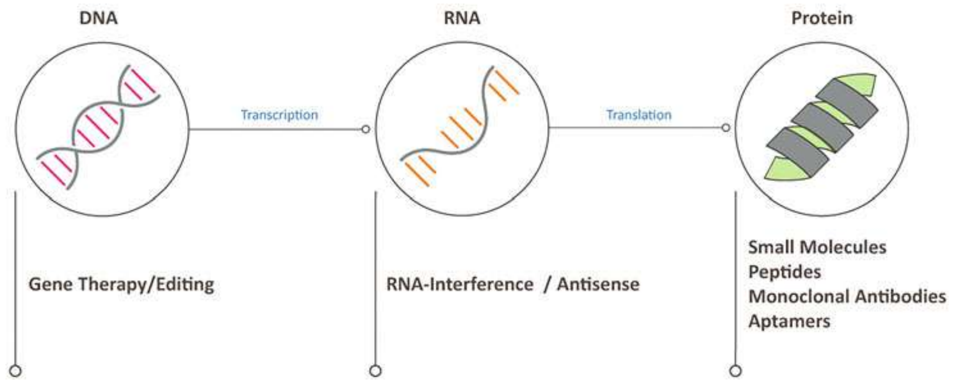


Fig. 8.1 The central dogma of molecular biology describes the transfer of information from DNA via RNA to the functional carrier proteins, which is determined by the order (sequence) of the respective monomers (nucleotides in the case of DNA and RNA; amino acids in the case of proteins). The function of the three biopolymers can be influenced by chemical compounds or by biochemical processes for chemical-biological experiments

8.1 RNA Interference

RNA interference (RNAi) is a biological process in which RNA molecules inhibit gene expression or translation by specifically neutralizing mRNA molecules.

Two types of small ribonucleic acid (RNA) molecules—microRNAs (miRNAs) and small interfering RNAs (siRNAs)—are central to RNA interference. RNAs are the direct products of genes, and miRNAs and siRNAs can direct enzyme complexes to specifically degrade mRNA molecules. This is also referred to as gene silencing.

The microRNA encoded in the genome is transcribed by RNA polymerase II/III and the so-called primary microRNA (pri-miRNA) is produced. This is cut by the nuclease Drosha while still in the cell nucleus, producing precursor miRNAs (pre-miRNA) that are 70–80 nucleotides long and organized as a double-stranded hairpin.

The pre-miRNA is transported out of the nucleus. In the cytosol, the pre-miRNA encounters the ribonuclease DICER. DICER forms a complex with the proteins PACT and TRBP, which cleaves the double-stranded pre-miRNA and unwinds it into a single strand. The resulting single strand is the miRNA that eventually binds to the Argonaute-2 protein. The complex of miRNA and Argonaute-2 protein binds to the mRNA in a sequence-specific manner predetermined by the miRNA. The Argonaute-2 protein recruits other proteins, such as GW182, PABP, CCR4-NOT, and PAN2 and 3, which ultimately form the so-called RNA-induced silencing complex (RISC).

The RISC-mediated gene silencing is carried out by three different mechanisms:

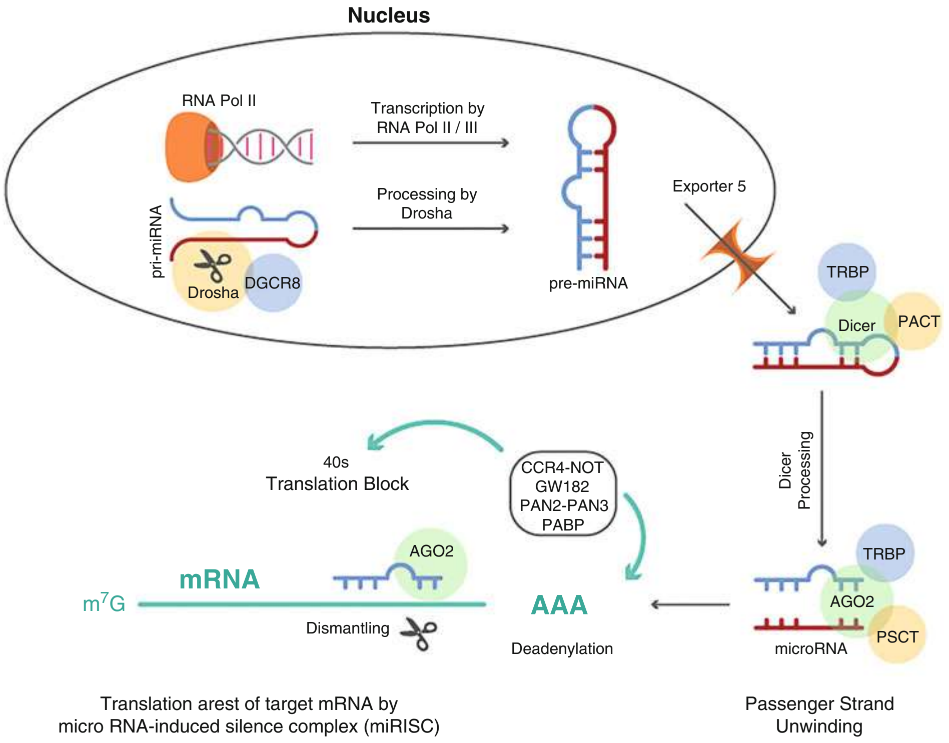


Fig. 8.2 Steps of gene muting. Abbreviations: AAA, poly-(A)-tail; AGO2, Argonaute2 protein; DGCR8, a microprocessor complex subunit; GW182, glycine-(g)-tryptophan-(w)-repeat-containing protein of 182 kDa; PABP, poly-(A)-binding protein; PACT, protein activator of interferon-induced protein kinase; PAN2 and PAN3, poly-(A)-nucleases 2 and 3; RNA pol, RNA polymerase; TRBP, HIV transactivating response RNA-binding protein-2

(i) The Argonaute protein is an RNase and catalyzes the cleavage-specific cleavage of target mRNA guided by bound miRNA.
(ii) GW182 inhibits translation initiation by preventing the formation of ribosomal complexes.
(iii) CCR4-NOT and PAN2-PAN3 facilitate deadenylation of the poly-(A) tail (AAA), leading to exonucleases degrading the mRNA.

With the help of small interfering RNAs (siRNAs), the mechanism can be used specifically from the outside to switch off a gene. The double-stranded siRNA is modeled on the pre-miRNA and is also cleaved by DICER and processed into a single strand. The resulting "artificial" miRNA binds to the Argonaute-2 protein, which in turn can degrade the target mRNA in a sequence-specific manner. siRNAs, therefore, represent an exciting drug concept, as in theory they can silence any gene (Fig. 8.2).

Summary

RNA interference offers the possibility to inhibit genes and gene products that cannot be addressed by established methods such as small molecule drugs or antibodies. By either introducing short double-stranded RNA (dsRNA) from the outside, targeted genes are switched off. The other way around, microRNAs can be manipulated, preventing gene silencing.

Further Reading

Jinek M, Doudna JA (2009) A three-dimensional view of the molecular machinery of RNA interference. Nature 457:405–412

9 RNA Interference in Drug Development

As previously mentioned, RNAi is very interesting for drug discovery. Ultimately, two concepts present themselves:

- Mimic miRNA by short double-stranded RNA (dsRNA) to induce gene silencing in a manner similar to RNAi.
- Inhibit a target miRNA and thereby block translational arrest.

For the first option, it is possible to add synthetic miRNAs or to develop siRNAs for the genes to be inhibited. For the second option, inhibitors (usually oligonucleotides) must be developed for the target miRNA.

9.1 Oligonucleotides as Chemical Probes: Antisense

It became apparent early on that small double-stranded oligonucleotide (small interfering RNAs) could specifically silence genes in cell experiments. Unfortunately, it soon became apparent that in practice both single-stranded and double-stranded DNA or RNA oligonucleotides (also known as antisense) are useless for clinical applications. DNA and RNA oligonucleotides are, on the one hand, not resistant to serum nucleases, and on the other hand, they are poorly celled permeable due to their negatively charged backbone. Consequently, they cannot enter the cell by diffusion. In addition, they often show off-target effects as a result of hybridization with other similar oligonucleotide sequences or activation of immune responses. Due to their pharmacokinetic profile, the general use of oligonucleotides in therapy is therefore limited. Only under certain circumstances have antisense agents been successful in clinical trials:

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_9

Chemical Modifications of Oligonucleotides

Simple oligonucleotides are unsuitable for therapeutic use because naked DNA or RNA units are not resistant to serum nucleases. Therefore, efforts have been made to develop oligonucleotide derivatives with increased lipophilicity compared to natural ribonucleic acids in the hope of overcoming inappropriate pharmacokinetic properties, such as lack of resistance to serum nucleases and cell permeability.

The most common chemical modification of oligonucleotides to improve resistance to nucleases and induce rapid and stable hybridization in vitro and in vivo is methylation of the 2′-OH group, termed 2′-O-methyl(2′-OMe) modification (Fig. 9.1). Accordingly, 2′-OMe-modified RNAs have been shown by independent studies to effectively block RISC. However, a disadvantage of 2′-OMe-modified anti-miRNAs is that they are still susceptible to degradation by serum exonucleases. Therefore, they are not suitable for in vivo applications. Serum exonucleases cleave phosphate bonds between nucleotides. There is, therefore, a need for modifications that induce resistance to serum exonucleases.

The most feasible synthesis approach is the replacement of a non-bridging oxygen atom in the phosphate backbone with a sulfur atom (sulfurization) to form a phosphorothioate (PS) bond. Consequently, their reported half-life in the bloodstream is 1 to 4 weeks. However, the enhanced stability of PS-containing oligonucleotides against serum exonucleases is accompanied by decreased hybridization affinity to their target miRNAs. However, the most remarkable aspect is that PS modification increases the binding affinity to plasma proteins. Subsequently, PS-containing oligonucleotides are effectively absorbed from the injection site into the bloodstream within 1–2 h compared to other modified antisense agents. Finally, their ability to bind to proteins is not limited to plasma proteins. Binding to tissue or cell surface proteins is much stronger. Therefore, PS-containing oligonucleotides show good uptake behavior in several tissues, such as the kidney, spleen, lymph nodes, adipocytes, bone marrow, and especially liver but not in skeletal muscle or brain. In summary, in addition to their resistance to serum exonucleases, PS-modified oligonucleotides exhibit unique pharmacokinetic behavior in that they can bind to proteins, which allows them to be distributed more efficiently in the body via the bloodstream and cellular uptake compared to other modified oligonucleotides. From there, PS modification can often be found in clinically tested anti-miRs.

Another nuclease-resistant chemical modification is the introduction of a methoxyethyl group at the 2′-sugar position (2′-MOE) analogous to the 2′-OMe modification (Fig. 9.1). In contrast to 2′-OMe, 2′-MOE-modified oligonucleotides show higher affinity for their target sequences due to their more lipophilic nature and, in addition, comparable serum nuclease resistance to PS-modified anti-miRs. In summary, 2′-MOE-modified antisense agents showed higher efficacy than those with 2′-OMe modification.

In contrast to 2′-MOE-modified anti-miRs, the introduction of a fluorine atom substituent at the C2′-ribose position (2′-F) does not affect serum nuclease resistance. However, the fluorine atom at the 2′-position induces the sugar ring into a high C3′-endoconformation characteristic of A-form duplexes, leading to an exceptional affinity for target RNAs (increased melting temperature Tm: up to 3 °C per nucleotide). More importantly, 2′-F-

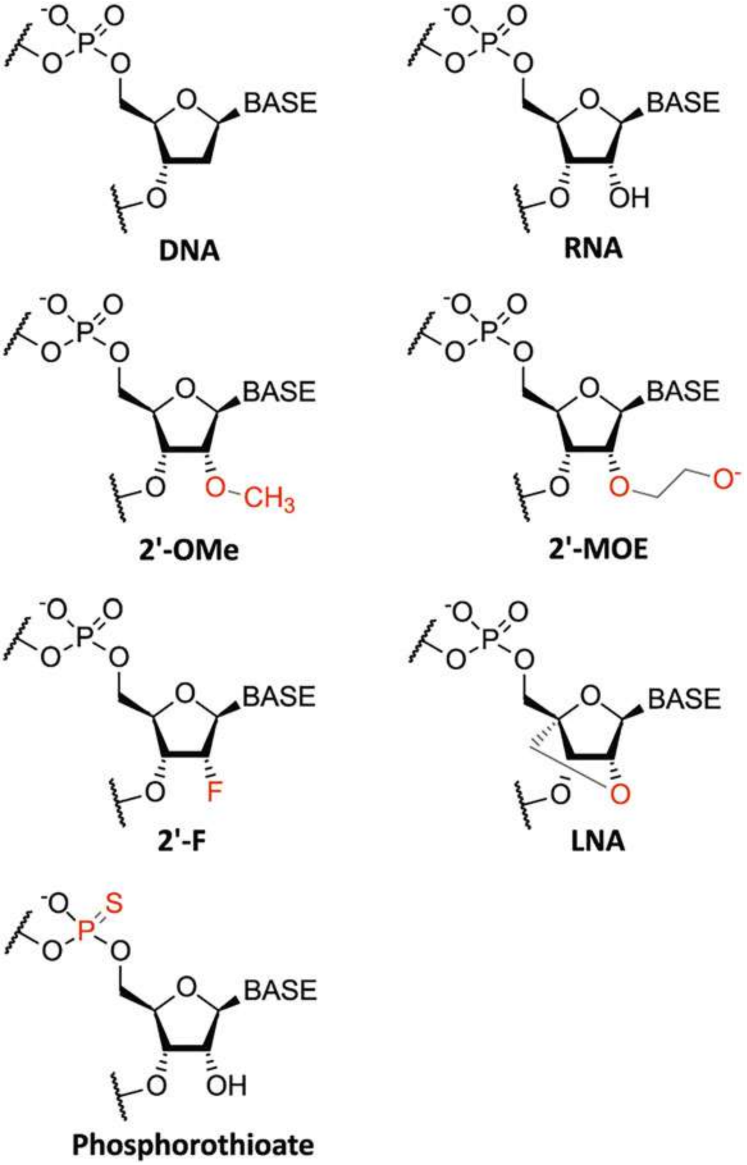


Fig. 9.1 Overview of chemical modifications used for the design of antisense oligonucleotides

modified anti-miRs were shown to promote RISC formation through protein recruitment, possibly due to their higher hybridization ability and/or preorganization in an A-form duplex.

However, the most successful nucleotide derivative is locked nucleic acid (LNA): Here, the ribose unit of an LNA nucleotide has an additional methylene bridge connecting the

2′-oxygen and the 4′-carbon. This “locks” the ribose in the C3 endo-confirmation, which enhances base stacking and backbone pre-organization in duplexes in A-form and protects against nucleases degradation. Therefore, the best-known anti-miR drug candidate is the LNA-PS-modified antisense oligonucleotide miravirsen (Fig. 9.2). The 15-mer miravirsen was developed by Santaris Pharma A/S (acquired by F. Hoffmann-La Roche AG in 2014 for US$450 million) and successfully tested in phase II clinical trials for the treatment of hepatitis C virus (HCV) infection. Miravirsen’s target is miRNA-122 (miR-122), which is expressed in the human liver. The miR-122 functions as an important host factor for the interaction of HCV with the 5′-untranslated region (5′UTR) of viral RNA by binding to two miR-122 sites in complex with Argonaute protein. The binding of the HCV genome to the miR-122-Argonaute complex protects viral RNA from nucleolytic degradation. Miravirsen blocks the miR-122-AGO complex, thereby inhibiting the essential interaction with HCV RNA for viral replication. Nevertheless, miravirsen is an exception in miRNA drug discovery. Due to its phosphorothioate modification, miravirsen accumulates in the liver after injection, where its target miR-122 is exclusively expressed. From there, the development of anti-miRs targeting miRNAs outside the liver is still a challenge.

In summary, various chemical modifications on ribose ring have been developed and are widely used in different combinations to obtain an anti-miR with the pharmacodynamic and pharmacokinetic parameters required for therapeutic use. To date, only the DNA/LNA-PS-modified anti-miRmiravirsen has reached the stage of phase III clinical trial, which, ironically, is exceptionally working. However, this raises the question of whether classical antisense agents are generally suitable for clinical use. It is likely that non-nucleotide modifiers hold promise for the development of more effective entities in miRNA drug discovery. One such example is N,N-diethyl-4-(4-nitronaphthalene-1-ylazo)phenylamine (ZEN): Behlke et al. from Integrated DNA Technologies, Inc. developed the ZEN monomer. They incorporated the ZEN monomer at both ends of a 2′-OMe-modified anti-miR, demonstrating that the ZEN modification increased potency and specificity compared to the unmodified oligonucleotide, while the modifications also showed low toxicity in cell culture.

Strategies for the Uptake of Oligonucleotides

The introduction of chemical modifications in anti-miR oligonucleotides effectively improved their resistance to serum nucleases. However, due to their negatively charged backbone, efficient delivery of antisense agents in vivo is still a challenge. Chemically modified anti-miR oligonucleotides are usually administered in high doses to have an effect and are therefore often directly absorbed by the liver and kidney, causing them to be rapidly excreted in the urine. Therefore, three strategies to improve the in vivo delivery of anti-miRs have been extensively investigated: The use of covalent conjugation with up-take enhancers, liposomes, and nanoparticles.

In 2004, Krützfeld et al. reported the use of 2′-OMe/PS-modified ribonucleic acids conjugated at their 3′-end with a cholesterol tag, so-called antagonistic miRNAs (antagomiRs), to enhance their cellular uptake. Although cholesterol tagging increases

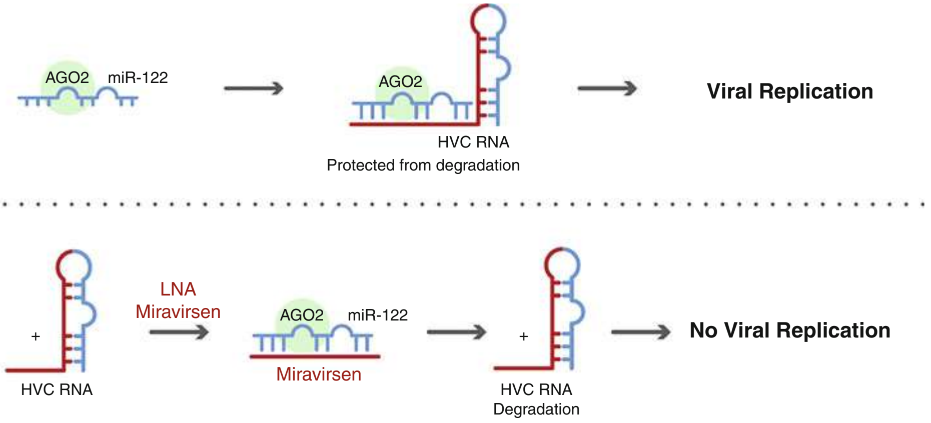


Fig. 9.2 Mechanism of the LNA-PS-modified oligonucleotide Miravirsen. (**a**) Hepatitis C virus RNA is either protected by AGO2 in the absence of Miravirsen or (**b**) degraded in the presence of Miravirsen as Miravirsen binds to the protecting AGO2 protein. (Illustration from Schmidt MF (2014) with permission)

cell permeability, the delivery of antagomiRs is still insufficient for therapeutic use because their doses required for in vivo administration are very high at 80 mg/kg.

More successful is the conjugation of an anti-miR—oligonucleotide with a trivalent N-acetylgalactosamine (GalNAc) carbohydrate tag. The trivalent GalNAc tag shows a high binding affinity to the asialoglycoprotein receptor (ASGPR). Originally, ASGPR is responsible for the efficient uptake of glycoproteins from serum via clathrin-mediated endocytosis. This mechanism was adapted for the efficient delivery of conjugated anti-miRs with remarkable in vivo activity with a mean effective dose of less than 0.05 mg/kg (ED50). The ASGP receptor is well conserved across species. From there, the GalNAc delivery strategy can be applied alongside human medicine.

An example of successful delivery of antisense agents is MRX34, a liposomal nanoparticle loaded with synthetic miRNA-34 mimetics developed by Mirnarx Therapeutics, Inc.: Human miRNA-34a is a down-regulated tumor suppressor in several cancers regulating more than 20 oncogenes, including the epidermal growth factor receptor (EGFR) pathway. The addition of artificial miRNA-34a to cancer cells has been shown to suppress oncogenes, and from there, it shows potential therapy. MRX34 has recently entered phase II clinical trials. Andreas Bader et al. recently reported that MRX34 was tested in combination with the EGFR tyrosine kinase inhibitor (EGFR-TKI) erlotinib against cancer cells, suggesting that several cancers previously unsuitable for erlotinib may be sensitive to the drug when combined with MRX34.

In summary, MRX34 represents two trends in miRNA drug discovery. On the one hand, it is delivered through an advanced formulation (liposomal nanoparticles) to increase the delivery of oligonucleotides into the cell. On the other hand, MRX34 is being tested in combination with reported drugs to identify synergistic effects.

Challenges of Antisense Agents in miRNA Drug Discovery

The main challenge of antisense agents in miRNA drug discovery remains their uptake from the injection site via the bloodstream and through the tissue into the cell. Admittedly, some tissues such as the liver and kidney are much more accessible to antisense agents than others. This explains the success of miravirsen and MRX34, both of whose targets are located in the liver. Even when anti-miRs reach their target tissues from the injection site, the most limiting step is cellular uptake and internal release of the anti-miR into the cytosol. For example, it was shown that only 1–2% of the nanoparticles entered the cell via clathrin-mediated endocytosis and macropinocytosis. This result indicates that firstly, previously existing uptake strategies such as liposomal nanoparticles are still inadequate for the delivery of anti-miRs into the cytosol and secondly, the clinically tested anti-miRs (miraviruses, MRX34, etc.) need to be administered at high doses to induce an effect. The administration of high doses always carries the risk of side effects.

Off-Target Effects

There are two types of off-target effects of anti-miRs: Hybridization-associated and hybridization-independent. Hybridization-associated off-target effects are due to the promiscuous binding behavior of an anti-miR to all family members of the miRNA target. This is because the target miRNA is bound to the Argonaute 2 protein as part of miRISC in the cell, and only the nucleotides at the 5′-end of the miRNA, called the seed region, are exposed to the solvent. Thus, the binding of anti-miR to its target miRNA occurs in a first, rate-dependent step to its seed region and only in a second, rapid step to the entire sequence. Since the seed region is largely the same for all family members, it is very difficult to target a specific family member with an anti-miR in vivo. Currently, the only solution for antisense agents in miRNA drug discovery seems to be careful selection of a miRNA target with a small number of family members or deliberate inhibition of a complete miRNA family.

In contrast, hybridization-independent off-target effects are associated with immune stimulation and in vivo toxicity. Anti-miRs, especially when administered in high doses, are recognized by the innate immune system. Cells of the innate immune system express Toll-like receptors (TLRs), a type of pattern recognition receptor (PRR), which recognize molecules of pathogens and therefore play a central role in microbial defense. Several TLRs (TLR3, TLR7, TLR8, and TLR9) have been identified that induce an interferon-alpha (IFN-α) response upon binding of a short nucleic acid, both single and double-stranded. For this reason, the interaction of a potential anti-miR drug is routinely studied in vitro and in vivo to avoid receptor activation. An interesting finding was that only the nine nucleotides at the 3′-end of the oligonucleotide that directly correspond to the seed region of miRNA are critical for TLR activation. Another sequence-independent off-target effect is in vivo toxicity. This is often recognized by inhibition of coagulation, activation of the complement cascade, and hepatotoxicity. Herein, PS-modified anti-miRs were shown to inhibit coagulation and transiently prolong clotting times in monkeys, likely due to their ability to bind blood plasma proteins. For example, hepatotoxicity is frequently observed with LNA-modified oligonucleotides, as indicated by organ and body weight loss in preclinical animal studies.

In conclusion, antisense—miRNA modifiers are only effective in certain circumstances such as liposomal nanoparticles in high doses with the risk of off-target side effects, in combination with other drugs or exceptionally in the case of miravirsen. However, their general use is hampered due to their poor delivery and uptake properties and off-target effects, probably due to their molecular size of more than 6000 Da and their negatively charged backbone. In addition, antisense oligonucleotides must be applied by injection, making treatment less convenient. In contrast, the success of small molecule drugs targeting proteins depends mainly on their oral application and their ability to reach targets in the cell by simple diffusion. It can, therefore, be questioned why the concept of protein-based drug discovery cannot be applied to small-molecule miRNAs.

Summary

Currently, the only possibility to use RNA interference in drug development is anti-sense oligonucleotides. Unfortunately, DNA and RNA molecules are not stable in the body. Therefore, chemical modification has been introduced to stabilize oligonucleotides for therapeutic use. Nevertheless, their widespread application is still very limited due to their lack of uptake into the cell. Better formulations such as nanoparticles should help. Nevertheless, only two drugs (patisiran and givosiran) based on siRNA are currently approved and one is under review (inclisiran).

Further Reading

Kurrek J (2009) RNA interference: from basic research to therapeutic applications. Angew Chem Int Ed Engl 48:1378–1398

Schmidt MF (2014) Drug target miRNAs: chances and challenges. Trends Biotechnol 32:578–585

Schmidt MF (2017) miRNA targeting drugs: the next blockbusters? Methods Mol Biol 1517:3–22

Part IV

Proteins

Peptides and Proteins 10

Peptides and proteins are functional biopolymers composed of 20 natural alpha-L-amino acids in varying order (sequence) (Fig. 10.1). Peptides are divided into oligopeptides with short amino acid chains and polypeptides with long amino acid chains. Long, naturally occurring polypeptides are called proteins or albumin. While DNA and RNA molecules primarily carry or transmit hereditary information, proteins have functions that allow them to perform tasks such as biocatalysis (enzymes) or signal transduction (receptors). In contrast to DNA and RNA, these abilities are based on the greater number and structural diversity of their building blocks, the amino acids.

10.1 Amino Acid → Peptide → Protein

The building blocks of peptides and proteins are amino acids. As the name suggests, they are chemical compounds with a basic amino and an acidic carboxyl group. The alpha-carbon atom carries four different substituents. Such a carbon atom is called a stereogenic center, chiral or asymmetric. The substituents can be arranged in two ways (Fig. 10.2). The two forms relate to each other like image and mirror image or like left and right hands. These are called enantiomers. These cannot be converted into each other by rotation. The naturally occurring (proteinogenic) amino acids are exclusively alpha-L-amino acids. Why only L-amino acids have prevailed evolutionarily to build the naturally occurring proteins is not known. The proteinogenic amino acids differ only in the residue attached to the alpha carbon.

Amino acids have a basic and an acidic functional group. They can, therefore, act as a base as well as an acid (Fig. 10.3). Depending on the pH of the environment, an amino acid can carry a positive charge, a positive and negative charge simultaneously (so-called Zwitterion), and only a negative charge. If the charge is positive and negative, the structure

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_10

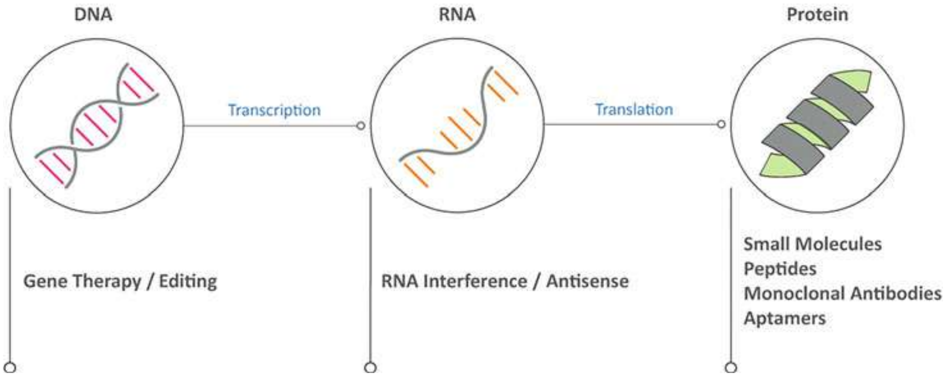


Fig. 10.1 The central dogma of molecular biology describes the transfer of information from DNA via RNA to the functional carrier proteins, which is determined by the order (sequence) of the respective monomers (nucleotides in the case of DNA and RNA; amino acids in the case of proteins). The function of the three biopolymers can be influenced by chemical compounds or by biochemical processes for chemical-biological experiments

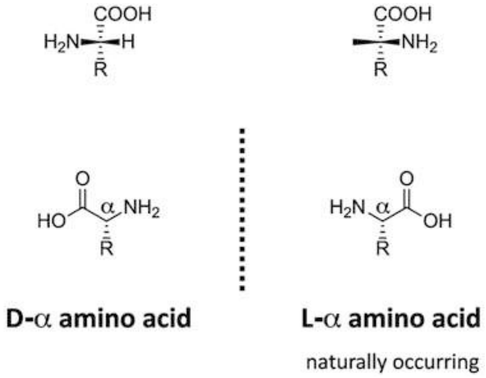
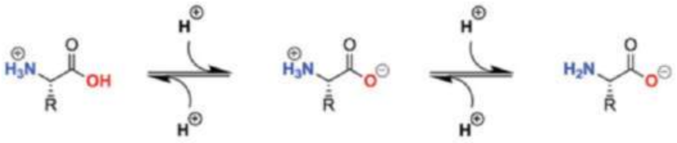


Fig. 10.2 Structure of the amino acid. The Fischer projection is used to represent the spatial structure of molecules in two dimensions. The horizontal ligands are facing the viewer, the vertical ligands are facing away from the viewer



$$K_s = \frac{c\,(H_3O^+) \cdot c\,(A^-)}{c\,(HA)} = K \cdot c\,(H_2O)$$

$$pK_s = -\lg\,(K_s \cdot 1\ \text{mol}^{-1})$$

Fig. 10.3 Zwitterionic structure of the amino acid

is said to be zwitterionic. The pH value at which the net charge of the amino acid is zero is called the isoelectric point (IEP).

The acid constant K_S indicates the strength of the acid. According to the law of mass action, to determine the position of the equilibrium of a reaction, the concentrations of the products are multiplied together and divided by the multiplied concentrations of the reactants. The calculated ratio is described using the equilibrium constant K. Since the concentration of water (c(H_2O)) remains constant in aqueous solutions, one includes c (H_2O) in the constant K. This results in the acid constant K_S with the unit mol/L. Usually, very small values are obtained. Therefore, analogous to the pH value, the negative decadic logarithm of $K_{S,}$ the so-called pK_a value, is formed. The following applies: The smaller the pK_a value, the stronger the acid. For example, the pK_a value of the very strong hydrochloric acid is −7 ($K_S = 10^{-7}$ mol/L), while the pK_a value of the weaker acetic acid is 4.74 ($K_a = 1.82 \cdot 10^{-5}$ mol/L).

As can be seen in Fig.10.4, the pK_a value for the carboxyl group of an amino acid is 3.6, which is slightly more acidic than acetic acid. The pK_a value for the amino group of an amino acid is 7.8. Accordingly, amino acids exist as zwitterions at the physiological pH of 7.35 to 7.45.

The 20 proteinogenic amino acids (Fig. 10.4) can be divided into four groups based on their side chains:

Overview

Non-polar Aliphatic Amino Acids

The side chains of the amino acids alanine, valine, leucine, isoleucine, and proline. Proline with its ring structure is called an imino acid because the amino group is not a primary amine but a secondary amine. As we will see, proline differs from the other amino acids in its structure and chemical reactivity. Glycine carries a second proton on the C_{alpha} atom instead of an aliphatic side chain, which is why it is the only amino acid that is not chiral. Despite these differences, it is simply added to this group.

Polar, Uncharged Amino Acids

The side chains of the amino acids serine, threonine, cysteine, methionine, asparagine, and glutamine are uncharged and hydrophilic and can form hydrogen bonds. Threonine is the only amino acid that has an additional asymmetric center at the second (beta) carbon atom. Serine, threonine, and cysteine can serve as nucleophiles in enzymatic catalysis. Cysteine, in particular, is a strong nucleophile because the proton of the thiol group can dissociate under alkaline conditions (pK_a = 9.1). Cysteines can be easily oxidized, leading to the formation of disulfide bonds between two cysteines, called cystines. Cystine disulfide bonds play a role in protein stability.

(continued)

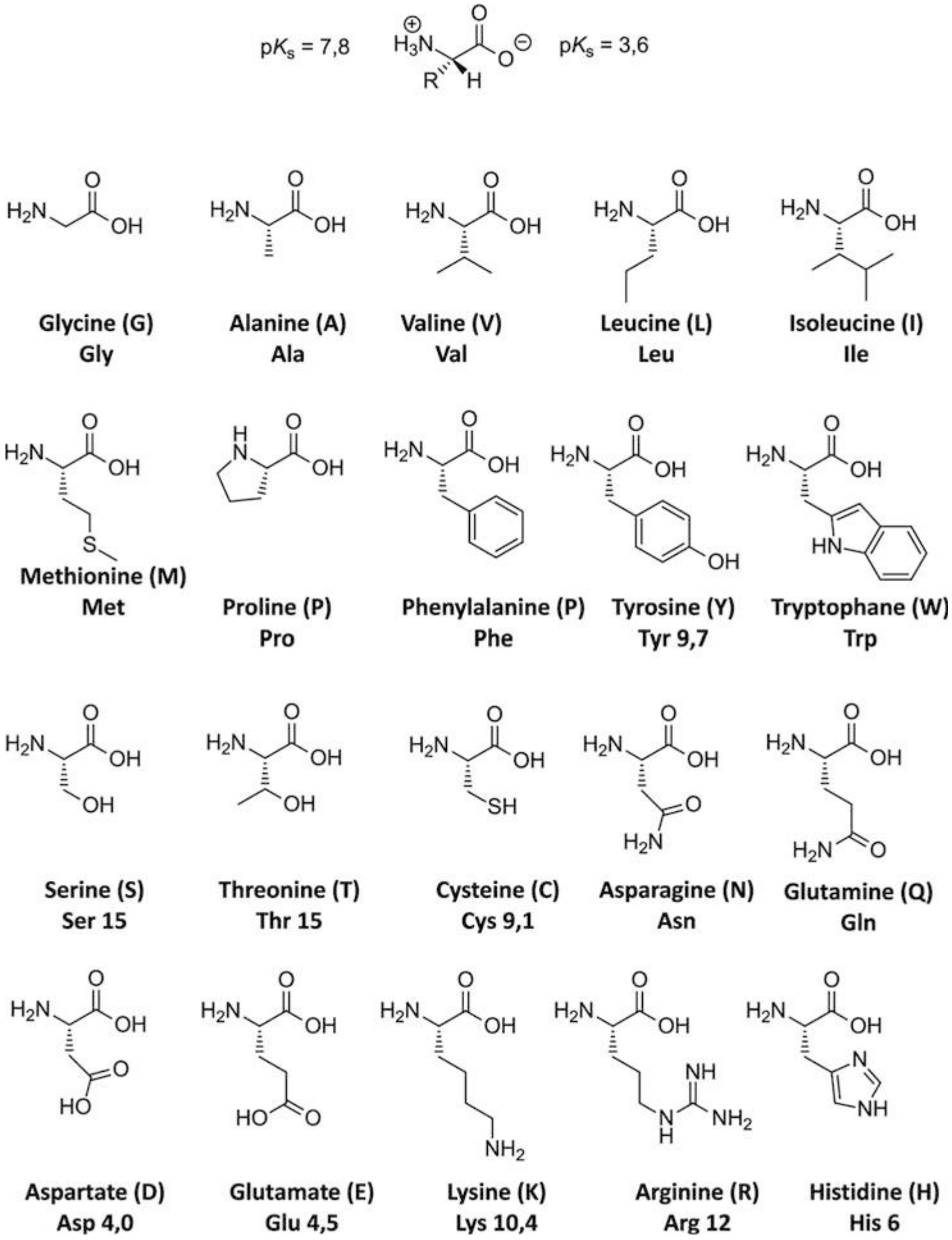


Fig. 10.4 Structures of naturally occurring alpha-L-amino acids

Aromatic Amino Acids

The amino acids tyrosine, phenylalanine, and tryptophan have aromatic side chains. These are hydrophobic. Nevertheless, the hydroxy group of tyrosine and the ring nitrogen of tryptophan can form hydrogen bonds, which have important functions in enzymatic catalysis.

Acidic (Negatively Charged) Amino Acids

Aspartic acid and glutamic acid also called aspartate and glutamate after their salts, have another carboxyl group. Their pK_a values are between 4 and 4.5, making them slightly more acidic than acetic acid. The acidic side chains can stabilize proteins through ionic bonds or are decisive factors in the acid or base catalysis of enzymes.

Basic (Positively Charged) Amino Acids

The side chains of the amino acids lysine, arginine, and histidine are basic. While lysine carries only one additional amino group, arginine has a guanidine and histidine an imidazole. The pK_a value of histidine is six and is therefore close to the physiological pH value. In enzymatic catalysis, histidine takes on the role of both a proton donor and a proton acceptor (buffer effect).

Peptides and proteins are polycondensed amino acids (polyamides). The amino reacts with the carboxyl group, splitting off a water molecule, to form an amide bond or the so-called peptide bond. The peptide chain formed has a polarity. One end is formed by an amino group (*N*-terminus), the other by a carboxylic acid (*C*-terminus). By definition, the amino acid is always indicated from the amino (left) to the carboxyl (right) end. The reason: In protein biosynthesis, the synthesis starts from the *N*-terminus (Fig. 10.5).

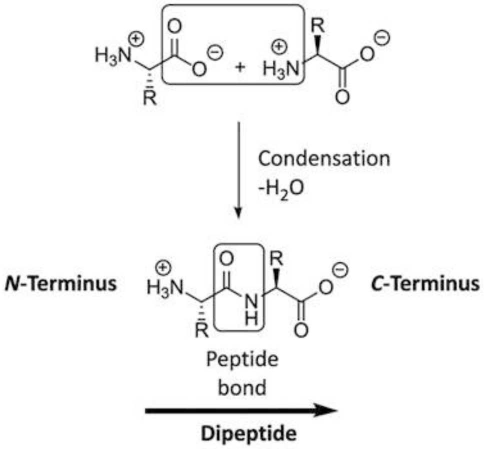


Fig. 10.5 Schematic peptide bond

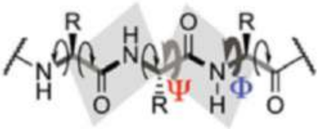


Fig. 10.6 The peptide bond is rigid and planar. Rotation angles in a peptide exist only to the left and right of the peptide bond

The structure of peptides and proteins is largely determined by the chemical property of the peptide bond. The notation of the peptide bond used in Fig. 10.5 represents only a boundary structure. As shown in Fig. 10.6, electrons distribute from the carbonyl to the amino group, forming a partial double bond. This limits the mobility of the peptide bond. As a result, the total of six atoms involved in the peptide bond lies in one plane. There is free rotatability around the single bonds of the C_{alpha} carbon atoms. If necessary, the rotatability can be influenced by the respective side chain.

Chain lengths of up to 20 amino acids are called oligopeptides. Up to 100 amino acids are called polypeptides or peptides in general. Molecules with more than 100 amino acids are called proteins or egg whites (due to their occurrence in the albumen of eggs). The classification is arbitrary, showing no relationship to functional or structural properties.

Four different structures are distinguished in proteins according to hierarchical sequence (Fig. 10.7):

Overview

Primary Structure

The primary structure is called the amino acid sequence.

Secondary Structure

Secondary structure is the spatial arrangement of an amino acid sequence.

Tertiary Structure

The tertiary structure represents the relative position of the secondary structure elements of a polypeptide chain in space to each other.

Quaternary Structure

The quaternary structure represents the relative position of several polypeptide chains in space to each other. The quaternary structure only occurs in proteins that consist of several polypeptide chains (subunits).

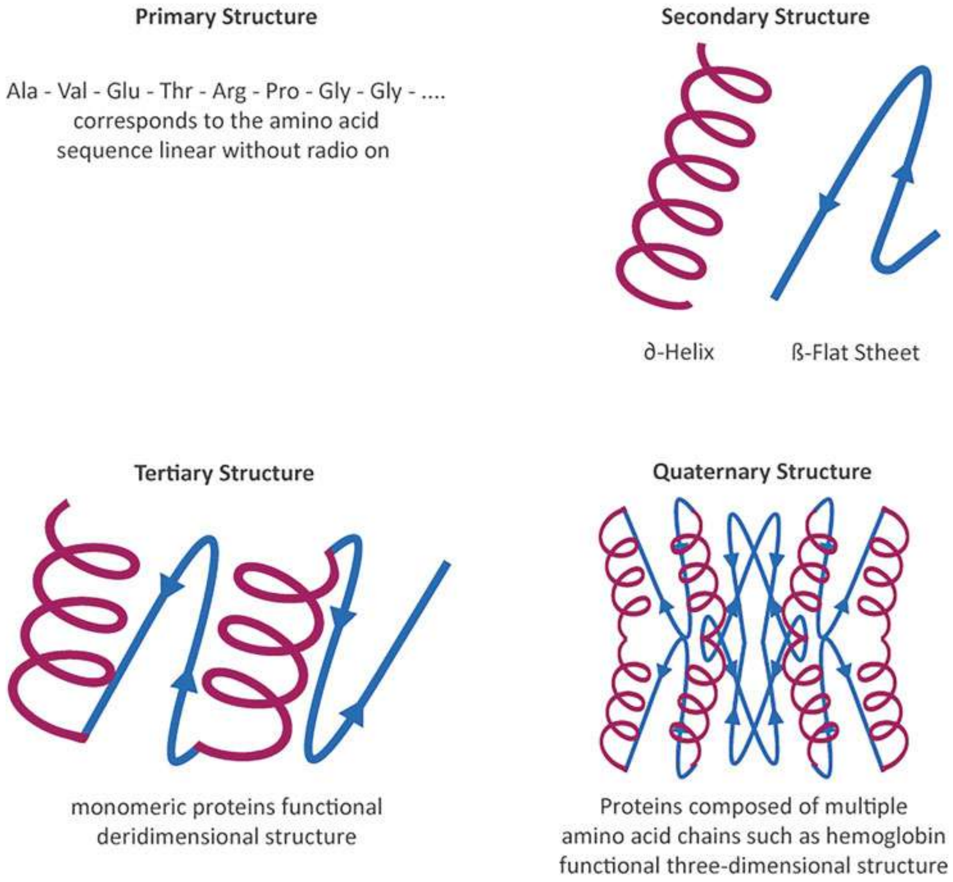

Primary Structure
Ala - Val - Glu - Thr - Arg - Pro - Gly - Gly - ....
corresponds to the amino acid
sequence linear without radio on
Secondary Structure
∂-Helix
ß-Flat Stheet
Tertiary Structure
monomeric proteins functional
deridimensional structure
Quaternary Structure
Proteins composed of multiple
amino acid chains such as hemoglobin
functional three-dimensional structure

Fig. 10.7 There are four levels of protein structure: Primary structure is the amino acid sequence. Secondary structure is the spatial arrangement of an amino acid sequence. In contrast, the tertiary structure represents the relative position of the secondary structure elements of a polypeptide chain in space. While the quaternary structure describes the relative position of several polypeptide chains in space to each other. The quaternary structure only occurs in proteins that consist of several polypeptide chains (subunits)

10.2 Peptide Synthesis

The synthesis of peptides has long been considered a challenge. The reasons for this are (Fig. 10.8)

- the reactive side chains of the amino acids,
- the complete linkage of the peptide bond,
- the stereoisomerism of the C_{alpha} atom.

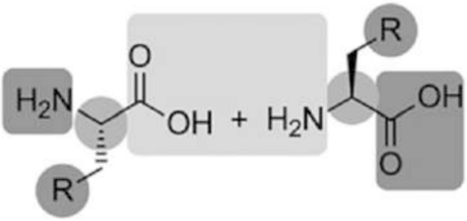


Fig. 10.8 Structure of two amino acids reacting to form a dipeptide (grey). The reactive side chain as well as the averted amino and carboxyl groups hinder a selective dipeptide synthesis (blue). In addition, two stereocenters exist at the C_{alpha} atoms, which do not have to be maintained during peptide binding (orange)

Sidechains

In view of reactive side chains of amino acids, which can lead to undesired side reactions, protective groups are used for them. A protecting group is a substituent that temporarily protects a particular functional group and therefore prevents an undesired reaction at that group. After the desired reaction has been carried out elsewhere in the molecule, the protecting group is cleaved off. Protective groups differ depending on the functional group as well as in their stability and the conditions of their cleavage.

Four different types of protecting groups have become established in peptide chemistry (Fig. 10.9): The acid-sensitive protecting groups, the benzyl-type and allyl-type protecting groups, and, in solid-phase chemistry, the fluorenylmethoxycarbonyl (Fmoc) protecting group (Fig. 10.10).

10.2.1 Orthogonality of Protection Groups

When using several protecting groups of different types, each protecting group can be cleaved individually and in any order due to the different cleavage reagents without attacking the other protecting groups. This is called orthogonality of protecting groups (Fig. 10.11). The synthesis of complex peptides is **not** possible without this strategy.

10.2.2 Peptide Bond Formation

After the amino acid is orthogonally protected and only the carboxyl group of one amino acid to be reacted and the amino group of the other amino acid are free, the condensation

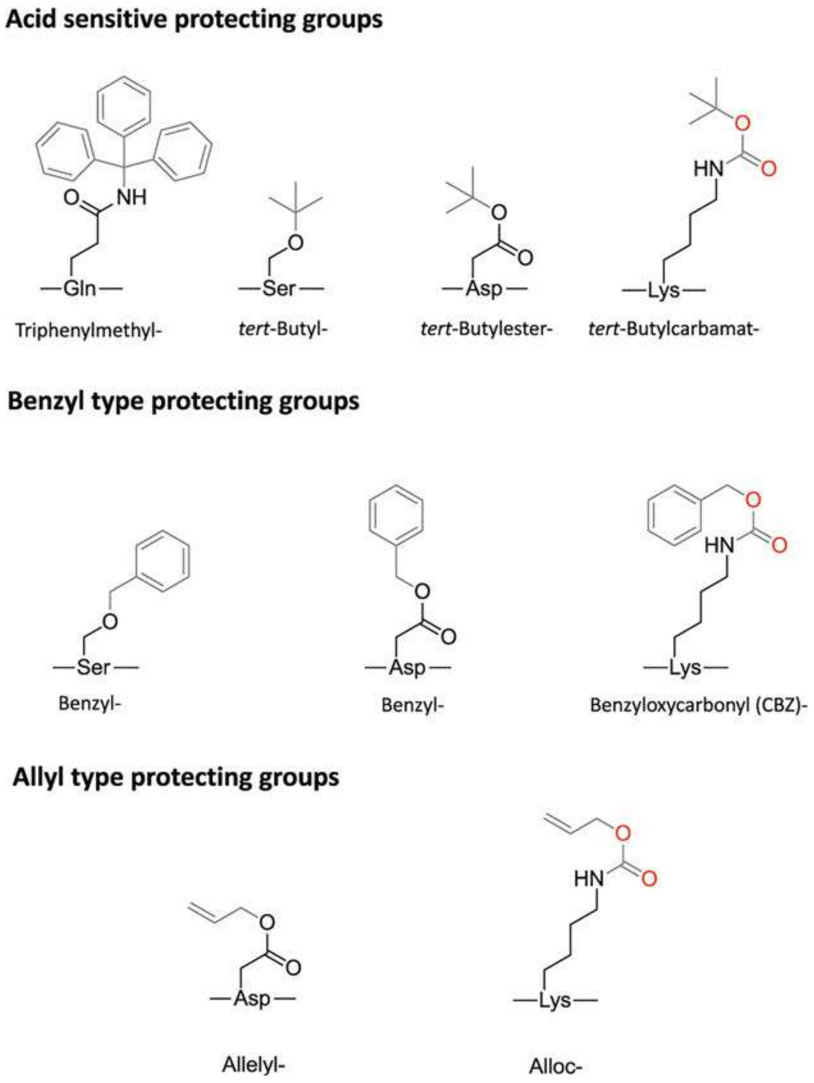


Fig. 10.9 Overview of protecting groups in peptide chemistry

reaction to the dipeptide should start by the simple application of heat. As shown in Fig. 10.12, this does not occur. Rather, the salt is formed.

To avoid the acid-base reaction, it is convenient to use a carboxylic acid derivative activated for a nucleophilic substitution reaction. Carboxylic acid chlorides are easy and inexpensive to prepare. However, it is found that the use of highly activated carboxylic acid

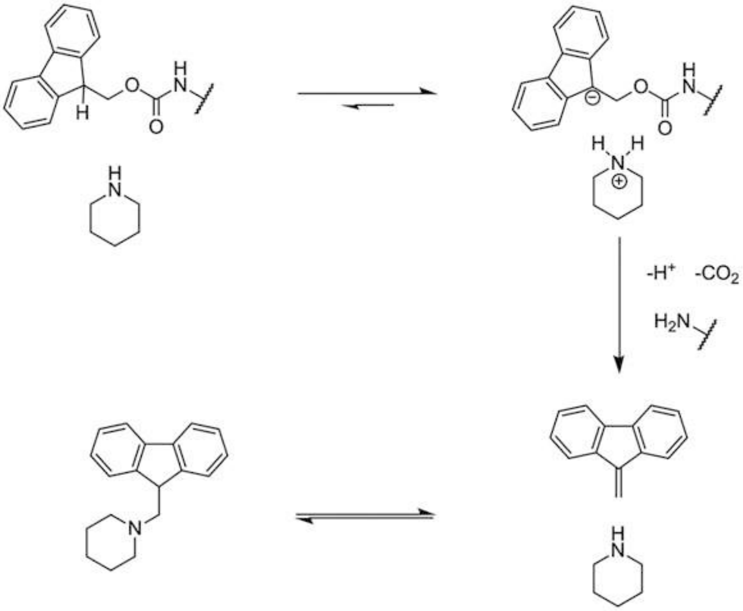


Fig. 10.10 The fluorenylmethoxycarbonyl (Fmoc) protecting group is used to protect the *N*-terminus in peptide solid-phase synthesis. The cleavage occurs with the help of the base piperidine. In addition to CO_2, the cleavage product is a UV-active adduct

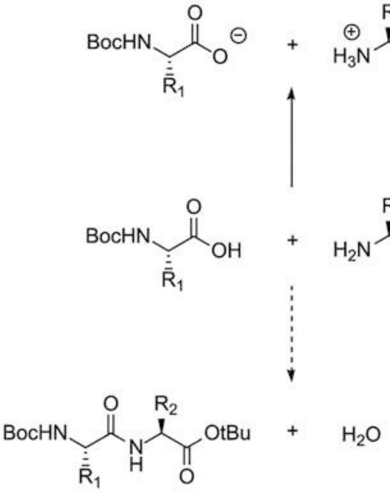


Fig. 10.11 Orthogonally protected L-lysine: Fmoc-protected *N-terminal* amino group (grey), with a *tert-butyl* ester-protected amino group of lysine (blue), and with a benzyl ester-protected *C*-terminus (orange)

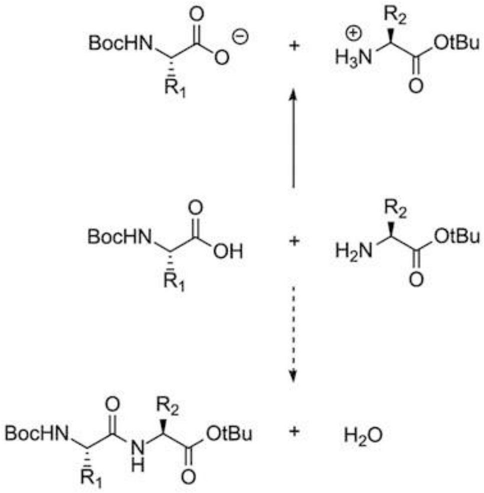


Fig. 10.12 Comparison of (**a**) attempted condensation reaction of two amino acids to form the dipeptide and (**b**) the actual formation of a salt with a carboxylate ion and an ammonium ion

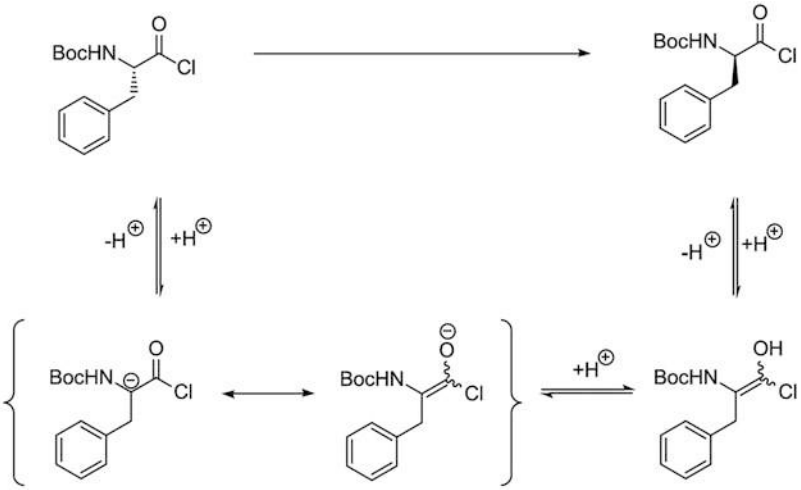


Fig. 10.13 Stereoisomerization of an *N*-protected, carboxylic acid chloride-activated amino acid via an enol intermediate

derivatives leads to a change in the stereoisomerism of the C_{alpha} atom, known as racemization. The chlorine atom of a carboxylic acid chloride is strongly electron-withdrawing, resulting in the formation of an enol or an oxazolone intermediate (Figs. 10.13 and 10.14), which as a consequence racemizes the stereocenter of the C_{alpha} atom.

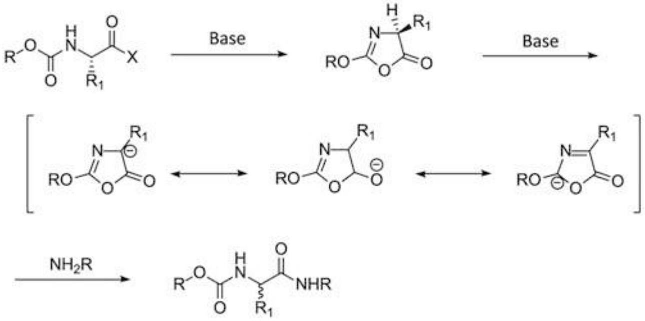


Fig. 10.14 Alternatively, racemization of the C_{alpha} atom can be accomplished by oxazolone formation. In this case, the protected amino group of the *N*-terminus attacks the activated carboxylic acid and forms a five-membered ring. This, in turn, can be attacked nucleophilically from both sides, whereby the isomerization occurs

To avoid racemization of the C_{alpha} atom, peptide chemistry uses less reactive carboxylic acid derivatives that activate sufficiently but not too strongly. In Fig.10.15, carboxylic acid derivatives are ordered according to their reactivity with a nucleophile. The rule for determining the reactivity of a carboxylic acid derivative with a nucleophile is the pK_a value of the acid of the leaving group corresponding to Brønsted. The leaving group of the carboxylic acid bromide is bromide. The corresponding acid is hydrogen bromide with a pK_a value of −9. The corresponding acid of the leaving group NH_3 of the carboxylic acid amide has a pK_a value of 23. In other words, a good leaving group must be able to stabilize a negative charge.

Here, *O*-acylisoureas, in particular, have proven to be suitable carboxylic acid derivatives for peptide chemistry. The activated amino acid is generated *in situ* by using carbodiimides, such as dicyclohexylcarbodiimide (DCC). As shown in Fig. 10.16, an *N*-terminal Boc-protected amino acid reacts with DCC to form an activated acylisourea. The *O*-acylisourea subsequently reacts with the *N*-terminus of the added amino acid to form a peptide bond, producing a dicyclohexylurea leaving group. The conversion of the carbodiimide bond to urea is the enthalpic driving force of this reaction.

However, the acylisourea has a disadvantage: It undergoes the side reaction of a 1,3 rearrangement, which leads to an unreactive *N*-acylurea. To prevent this, it has become established practice to add 1-hydroxybenzotriazole (HOBt), which reacts with the *O*-acylisourea to form a weaker activated ester before the 1,3 rearrangement reaction can occur. However, HOBt is explosive and has therefore been banned in some countries, such as the UK, in recent years as a result of the war on terrorism.

Another advantage of DCC is that the end product, dicyclohexylurea, is poorly soluble and fills in common solvents, and can be easily filtered off. Therefore, DCC is not used in peptidesolid-phase chemistry. Instead, diisopropyl carbodiimide (DIC) and 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) are used.

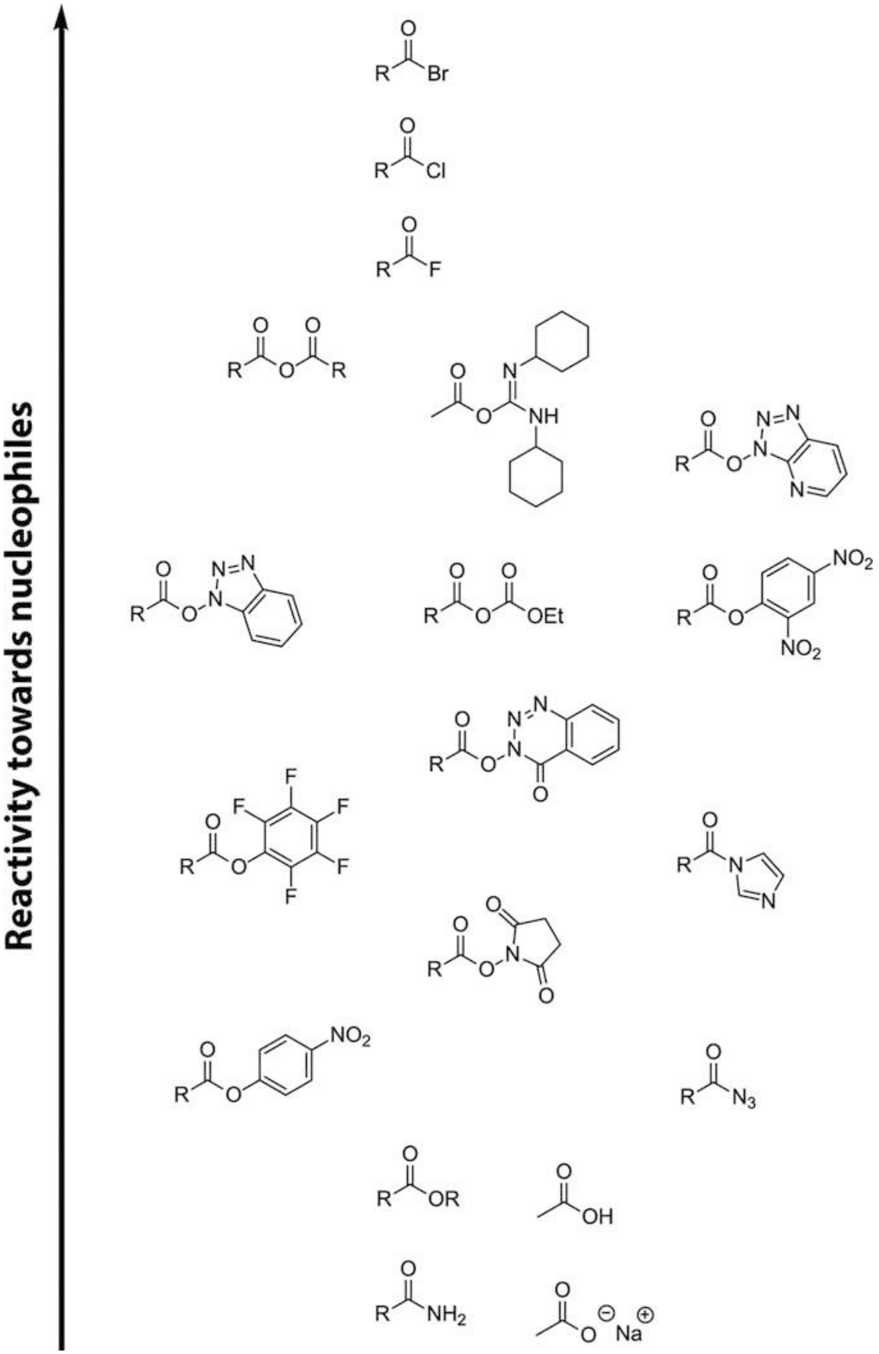


Fig. 10.15 The reactivity of carboxylic acid derivatives towards nucleophiles is largely determined by the property of the leaving group to be able to stabilize a negative charge

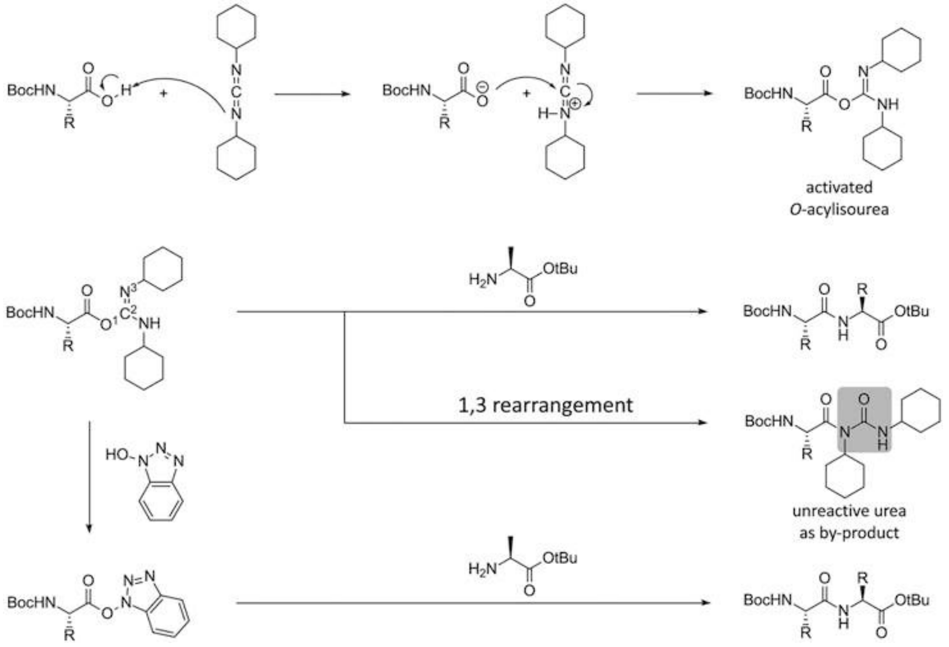


Fig. 10.16 The amino acid reacts with DCC to form the activated *O*-acylisourea. This may react with the amino acid to be coupled or undergo a 1,3 rearrangement to an unreactive *N*-acylureaas a side reaction. To prevent the latter, the addition of HOBt allows weaker activated esters to form, which subsequently react with the *N*-terminus of the amino acid to be coupled

In addition to carbodiimides, aminium salts such as HBTU, HATU, TBTU, and phosphonium salts such as TFFH, PyBrOP, BOP, or PyBOB are also used. HOBt, HOAt, CDI, DSC, or HOSu are used as additives for carbodiimide coupling (Fig. 10.17).

10.3 Solid-Phase Peptide Synthesis

The synthesis of longer peptides in solution is extremely laborious, as after each coupling step the product has to be laboriously purified in order to start the next one. The *solid-phase peptide synthesis* (SPPS) method developed by Robert Merrifield allows the rapid assembly of the peptide chain by successive reactions of amino acids on an insoluble carrier polymer (Fig. 10.18).

The core of solid-phase peptide synthesis is the solid support, usually, polysterol resin beads functionalized with reactive groups such as amine or hydroxy groups (Fig. 10.19). The first amino acid is coupled to this. The growing polypeptide chain is covalently bonded to the support throughout the synthesis. Excess reagents and by-products are removed by

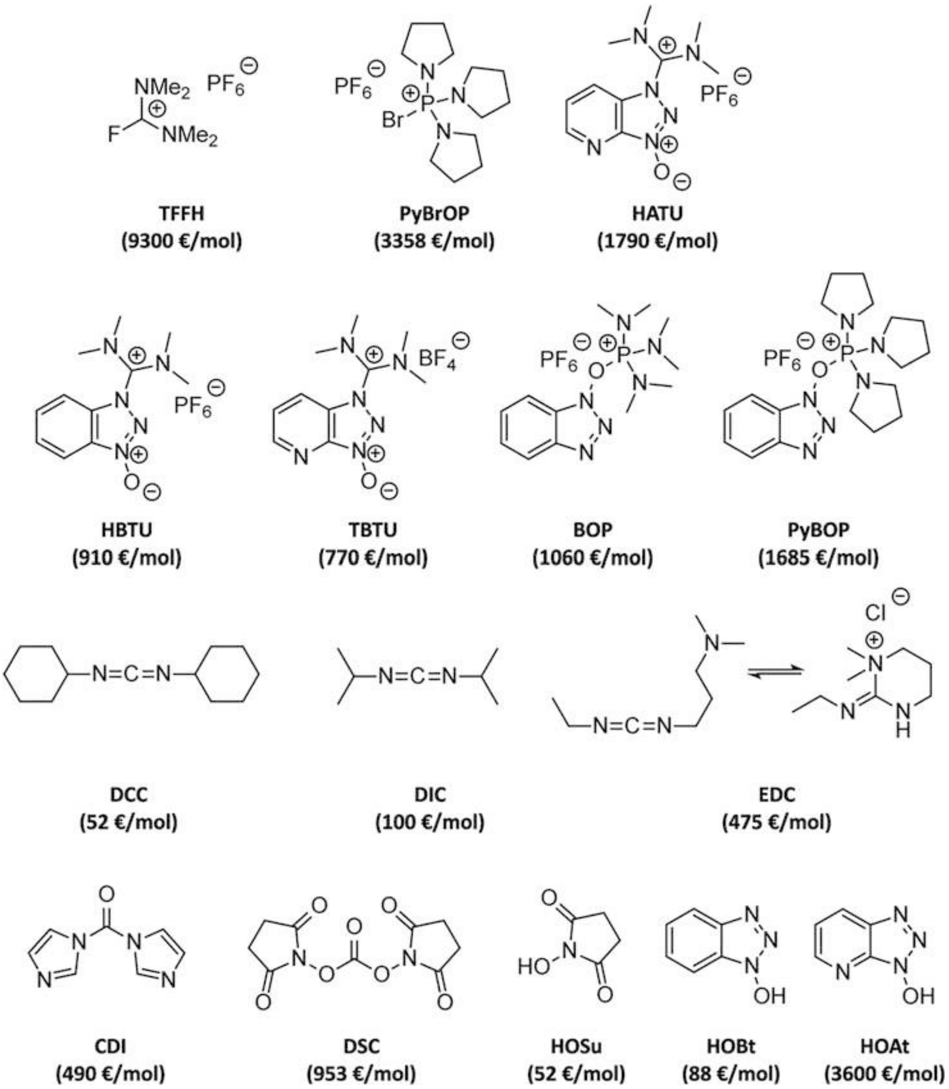


Fig. 10.17 Structures and prices of coupling reagents used in peptide synthesis

washing and filtration. This eliminates the time-consuming isolation of the product peptide after each reaction step when synthesizing the polypeptide in solution.

Analogous to peptide synthesis in solution, the amino acids used are orthogonally protected. This means that the *C*-terminus of the amino acids used is exposed, whereas the *N*-terminus is provided with protective groups that cannot be deprotected simultaneously with the protective groups of the side chains. Compared to protein biosynthesis, the polypeptide chain grows from the *C*-terminal amino acid to the *N*-terminal amino acids

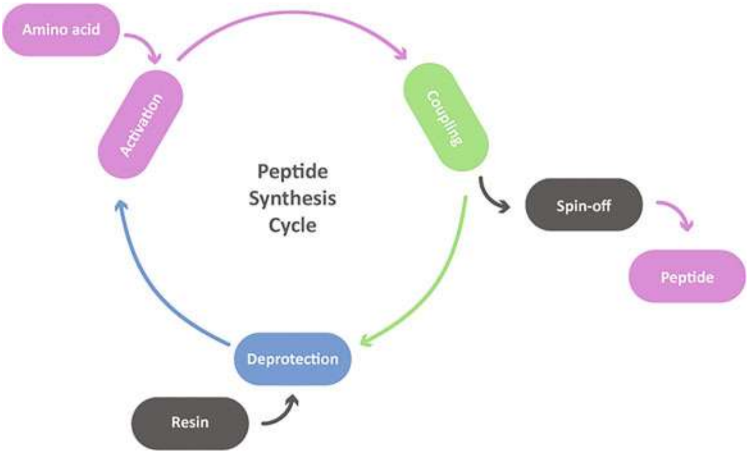


Fig. 10.18 Principle of peptide solid-phase synthesis using the example of a peptide synthesis cycle. A protected amino acid is bound to a solid phase, also called resin. This is deprotected (blue), and then an activated amino acid is added (red), which couples to the immobilized amino acid. Subsequently, the newly immobilized amino acid is also deprotected and the synthesis cycle starts again from the beginning until the desired sequence is present and the peptide is cleaved from the solid phase (grey)

and thus in the opposite direction compared to nature. Since in nature peptides and proteins start with the *N-terminus*, this convention in writing that the *N-terminus* is always on the left has prevailed despite reverse synthesis.

But why are peptides artificially synthesized from the *C*-terminus contrary to the model from nature? The reason is that by immobilizing the *N*-terminus and activating the carboxyl group of the amino acid to be coupled, one generates an excess. This excess leads to rapid acylation of the *N*-terminus. More importantly, the reaction must be 100%, otherwise, the free amino groups will react with another amino acid in the next synthesis cycle. This results in peptides that differ in length and sequence.

The reverse strategy would also mean that the carboxyl group must be activated on the resin. This is usually complicated by immobilization due to steric hindrance. As a consequence, some carboxylic acids would not be activated and could not react with the free amino group of the amino acid to be coupled.

The general principle of peptidesolid-phase synthesis is the alternating sequence of *N*-terminal deprotection and coupling reactions. The resin is washed between each of these steps. As shown in Fig. 10.18, an amino acid is attached to the resin. This is followed by deprotection of the amine. Acylation of the *N*-terminus is then carried out by the addition of an activated amino acid in excess. The resin is then washed again to remove the remaining unbound amino acid. This cycle is repeated as desired until the desired sequence is obtained. This is followed by cleavage from the solid phase and deprotection of the side chains. The crude peptide is precipitated in diethyl ether to remove residual organic soluble

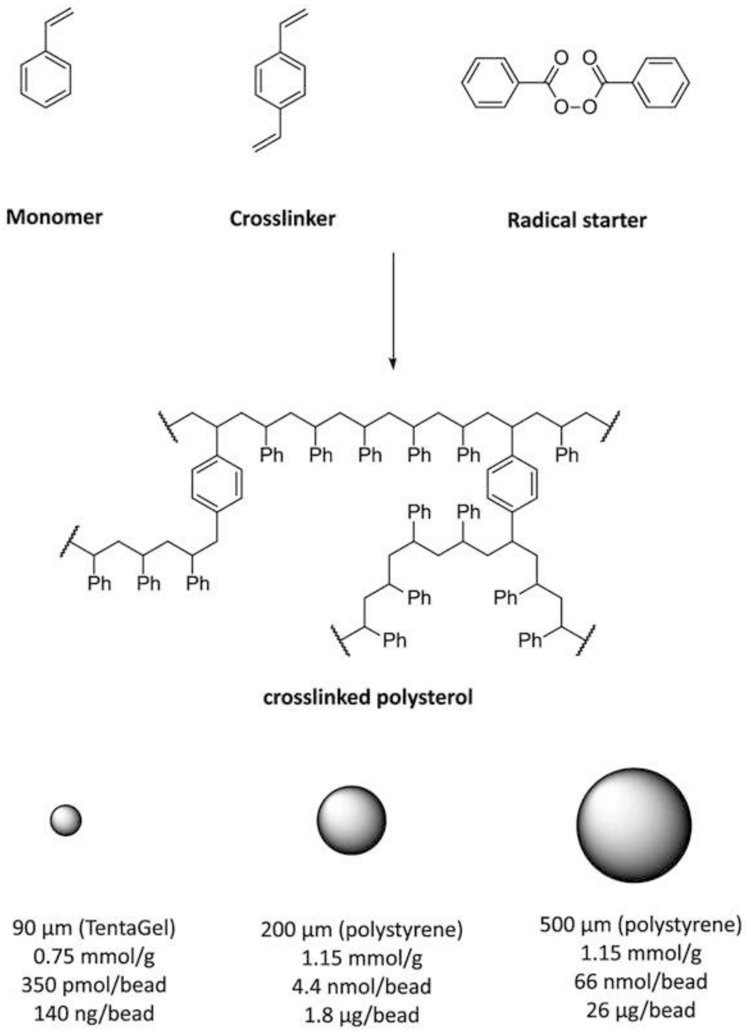


Fig. 10.19 Synthesis and structure of the polysterol resin beads used in peptide solid-phase synthesis

by-products. Further purification can be performed by HPLC. However, this is usually unnecessary for shorter peptides.

Unfortunately, peptide synthesis at the solid phase is not possible indefinitely. If the peptide sequence becomes longer, non-linear structures increasingly form, which prevent access to the free amino groups to be acylated. As a result, peptide mixtures are formed that are almost impossible to purify. Therefore, other methods are used for the synthesis of longer peptides or proteins.

10.3.1 Merrifield Synthesis

The original method for peptide synthesis, developed by Robert Merrifield, is based on the *tert*-butyloxycarbonyl (Boc) protecting group as a temporary *N*-terminal α-amino protecting group. The Boc protecting group is removed with the acid trifluoroacetic acid (TFA) (Fig. 10.20).

The side-chain protecting groups used in the Merrifield process are typically benzyl or phenyl-based. The cleavage of the peptide from the solid support occurs simultaneously with the removal of the side chains using anhydrous hydrogen fluoride by hydrolytic cleavage. A disadvantage of this approach is the use of hydrogen fluoride, which has a strong corrosive effect on skin and mucous membranes.

The Merrifield method is now considered obsolete, in particular, due to the use of hydrogen fluoride (hydrofluoric acid). Nevertheless, it is still used today in the synthesis of peptides with base-sensitive moieties such as depsipeptides, since in the predominantly used Fmoc-SPPS the deprotection step is carried out by treatment with a base.

10.3.2 Fmoc SPPS

As stated earlier, the Fmoc strategy has become widely accepted in solid-phase peptide synthesis. The use of the *N-terminal* Fmoc protecting group allows a milder deprotection scheme than in the Merrifield method. The base piperidine (20%) in DMF is typically used for the deprotection of Fmoc (Fig. 10.21). On the one hand, the free amine is neutral and consequently, no neutralization of the peptide resin is required as in the case of the Merrifield process. On the other hand, the released fluorenyl group is a chromophore. The Fmoc deprotection can be monitored by UV absorption of the reaction mixture, which consequently allows the calculation of the bound amino acids to the solid phase.

In addition, the Fmoc group can be cleaved under mild basic conditions while being stable to acids. This allows the use of orthogonal side-chain protecting groups such as Boc and tBu, which can be removed under milder acidic cleavage conditions, such as trifluoroacetic acid. Therefore, as shown in Fig.10.22, acid-labile linkers have become established in the Fmoc-SPPS. Sidechain deprotection and cleavage from the resin are carried out in one step by the addition of trifluoroacetic acid. However, a less soluble TFA salt is obtained as the crude peptide compared to the fluoride salt obtained in the Merrifield process.

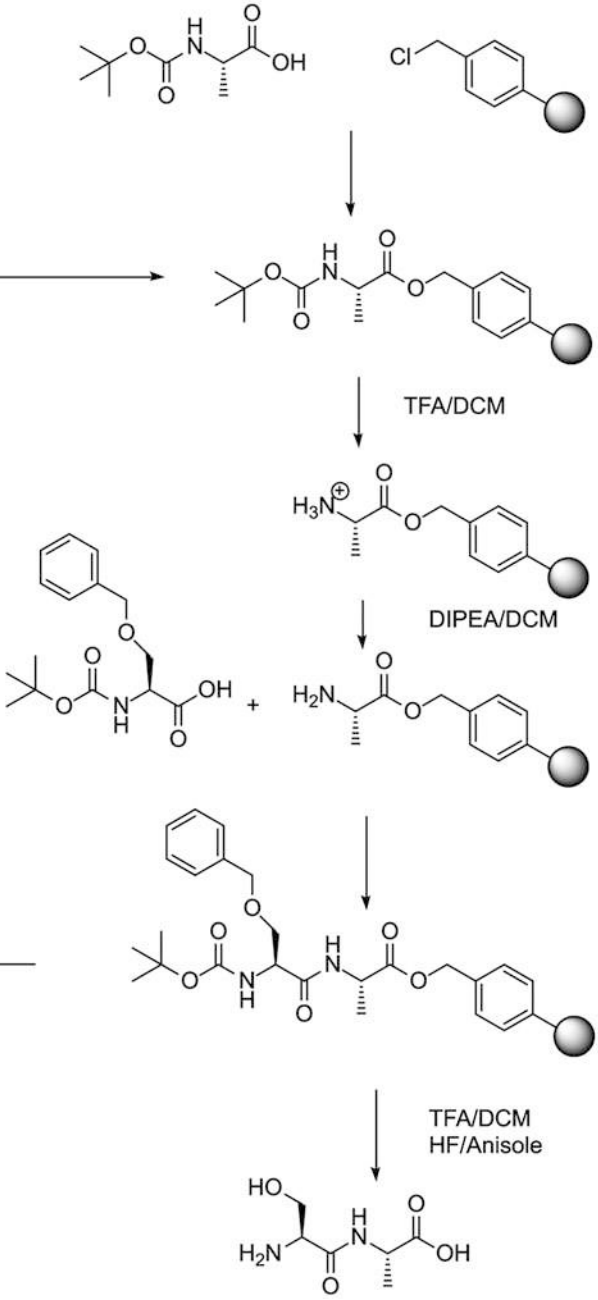


Fig. 10.20 Reaction course of peptide synthesis according to Merrifield

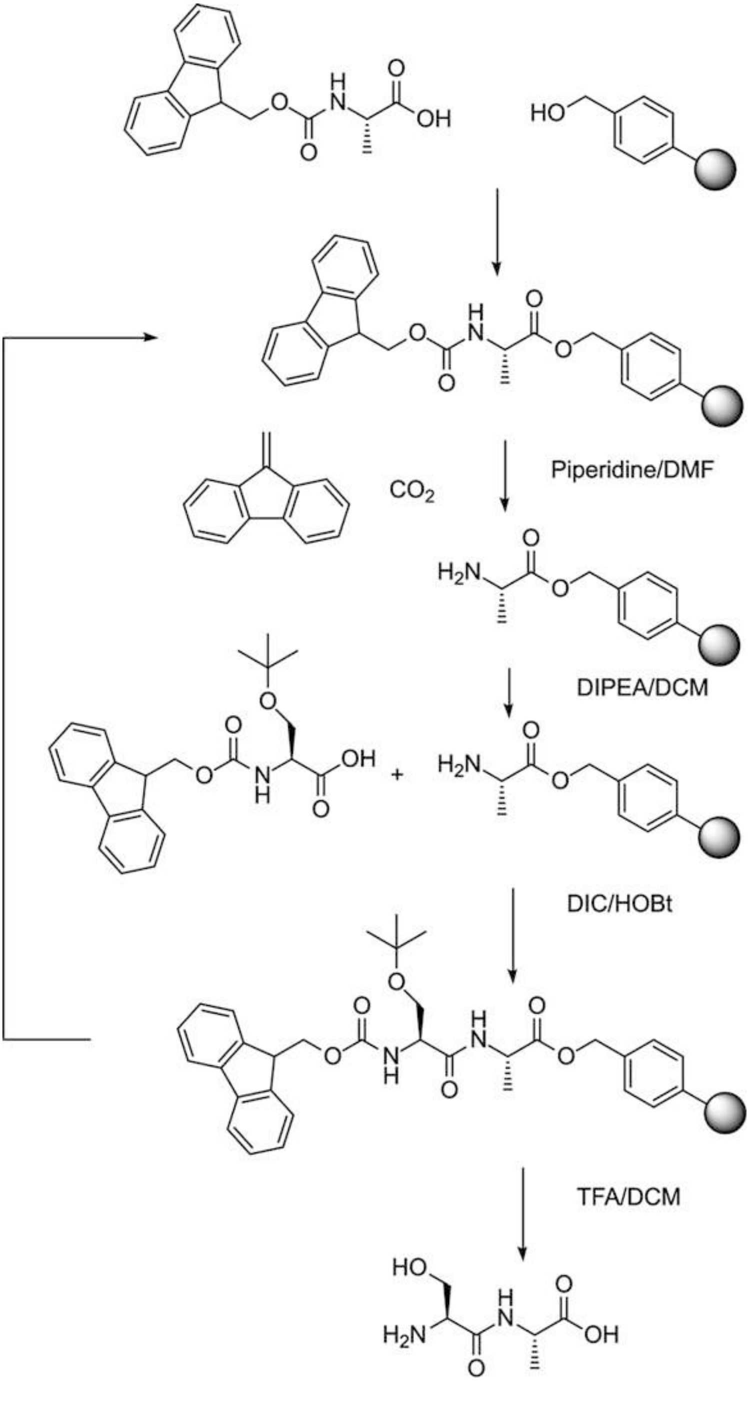


Fig. 10.21 Reaction course of peptide synthesis in the Fmoc-SPPS

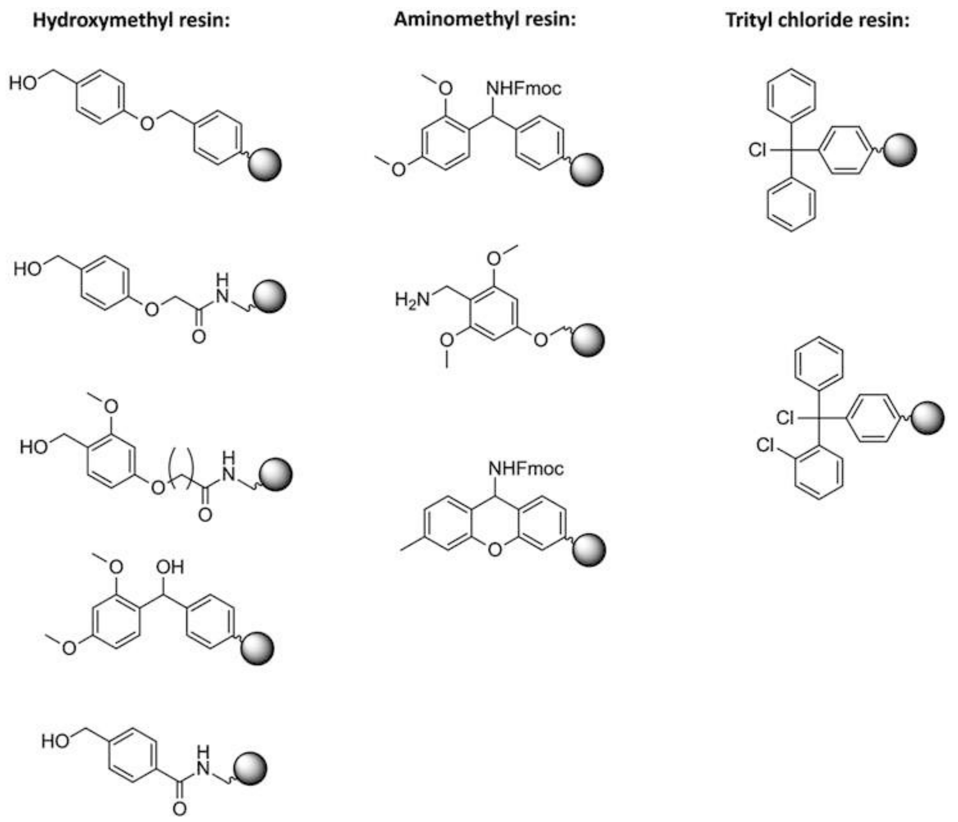


Fig. 10.22 Three main types of linkers are used in peptidesolid-phase synthesis: Hydroxymethyl, aminomethyl, and trityl chloride linkers

General Protocol of Fmoc Solid-Phase Synthesis

i. **Occupancy of 2-chlorotriphenylmethyl chloride resin with Fmoc-amino acids**

The resin used (loading 1.6 mmol/g) (1 eq.) was washed twice with DMF, swollen with DCM for 10 min, and then aspirated. Two equivalents were suspended in a mixture of dry DCM and dry DMF (1:1), mixed with six equivalents of DIPEA, added to the resin, and allowed to react on the shaker for 1 h. The resin was first washed three times with DMF and then three times with DCM, dried in vacuo, and the degree of occupancy was elicited by Fmoc determination. If the degree of occupancy was too low ($B < 0.4$ mmol/g), the resin was coated again according to the above instructions. Following the loading, unconverted chloride functions were saturated by reaction with 10 equivalents of MeOH for 5–10 min (“capping“) (Table 10.1).

(continued)

Table 10.1 Overview of the protected amino acids used in the Fmoc-SPPS with their molecular weight

Name amino acid	Protected amino acid	Molecular weight (g/mol)
Alanine	Fmoc-Ala-OH	311.3
Arginine	Fmoc-Arg(Pbf)-OH	648.8
Asparagine	Fmoc-Asn(Trt)-OH	596.7
Aspartic acid	Fmoc-Asp(OtBu)-OH	411.5
Cysteine	Fmoc-Cys(Trt)-OH	585.7
Glutamine	Fmoc-Gln(Trt)-OH	610.7
Glutamic acid	Fmoc-Glu(OtBu)-OH	425.5
Glycine	Fmoc-Gly-OH	297.3
Histidine	Fmoc-His(Trt)-OH	619.7
Isoleucine	Fmoc-Ile-OH	353.4
Leucine	Fmoc-Leu-OH	353.4
Lysine	Fmoc-Lys(Boc)-OH	468.5
Methionine	Fmoc-Met-OH	371.5
Phenylalanine	Fmoc-Phe-OH	387.4
Proline	Fmoc-Pro-OH	337.4
Serine	Fmoc-Ser(tBu)-OH	383.4
Threonine	Fmoc-Thr(tBu)-OH	397.5
Tryptophan	Fmoc-Trp(Boc)-OH	526.6
Tyrosine	Fmoc-Tyr(tBu)-OH	459.6
Valine	Fmoc-Val-OH	339.4

ii. Fluorenylmethoxycarbonyl (Fmoc) deprotection

After successful loading of the resin, the base-labile 9-fluorenylmethoxycarbonyl (Fmoc) protecting group was cleaved off using 20% piperidine/DMF solution (v/v). The cleavage was carried out by adding the piperidine reagent twice for 5 min each, after which the resin was washed first three times with DMF and then three times with DCM.

iii. Peptide coupling protocol

The resin containing the *N*-terminal deprotected occupied amino acid (1 eq.) was swollen in dry DMF for 10 min. Meanwhile, the *N*-terminal protected amino acid to be coupled (5 eq.) was dissolved with the activation reagent 1-hydroxybenzotriazole (HOBt) (5 eq.) in the required amount of dry DMF. After 5 min, the coupling reagent *N,N′*-diisopropyl carbodiimide(DIC) (5 eq.) was added. During the 10 min activation

(continued)

time that now followed, the resin was sucked dry. Subsequently, the mixture of Fmoc amino acids, HOBt, and DIC dissolved in DMF was added to the resin and shaken for 3 h. After this reaction time, the resin was sucked dry and washed first three times with DMF and then three times with DCM. Steps ii and iii were repeated as desired.

iv. **Acetylation of the *N-terminus***

After successful peptide coupling, the *N*-terminal Fmoc protecting group was cleaved off to ii and the now free *N*-terminus was acetylated by the addition of five equivalents of acetic anhydride. The reaction was carried out twice for about 20 min.

v. **Peptide cleavage and purification**

The resin was suspended in 95% TFA/water (v/v) (approximately 2 mL per 100 mg resin). In the case of Trt-protected side chains, a cleavage solution of 95% TFA, 2.5% H_2O, and 2.5% triisopropylsilane was used as a scavenger of the trityl cation. In the presence of sulfur-containing side chains, cleavage was carried out using 92.5% TFA, 2.5% H_2O, 2.5% triisopropylsilane, and 2.5% ethanediol or optionally thioanisole. The general addition of water was to remove any trifluoroacetic anhydride. After a reaction time of 3 h, the solution was filtered off.

The splitting solution was diluted with dist. Water and concentrated on the rotary evaporator. Subsequently, the dry matter was resuspended with ice-cooled diethyl ether. In the process, the peptide precipitated. To improve precipitation, the mixture was placed in the ice cooler for 20 min. The suspension was then centrifuged and the supernatant solution decanted. The remaining pellet was washed twice more with ice-cooled ether (2 mL each) in an ultrasonic bath, centrifuged and the supernatant solution decanted.

Subsequently, the peptide was taken up in water (in case of solubility problems: Addition of acetonitrile or dioxane), frozen and lyophilized.

vi. **UV/Vis spectroscopy**

UV/Vis and fluorescence spectra were performed on the Lambda 2 UV/Vis spectrometer (Jasco Germany) using Suprasil 0.1 (cm) quartz cuvettes.

To determine the resin occupancy of Fmoc amino acids, about 5 mg of dry resin was accurately weighed into a 10 ml volumetric flask, made up with 20% piperidine in DMF, and incubated for 90 min at RT. In the recorded UV spectrum of the cleavage solution, the degree of occupancy of the resin was determined based on the three absorption maxima of the *N*-(9-fluorenylmethyl)-piperidine adduct at 267 nm, 289 nm, and 301 nm. The true loading B was taken as the average of the three

(continued)

maxima. Taking into account the conversion factors and cuvette dimensions, the loading B is obtained by Labert-Beer's law to:

$$B = \frac{100000 \cdot E_\lambda}{\varepsilon_\lambda \cdot m_{(Resin\ weigh-in)}} \, [\mathrm{mmol/g}]$$

Extinction coefficients:
$\varepsilon_{267\mathrm{nm}} = 17{,}500\ \mathrm{cm}^2/\mathrm{mol}$, $\varepsilon_{289\mathrm{nm}} = 5800\ \mathrm{cm}^2/\mathrm{mol}$, $\varepsilon_{301\mathrm{nm}} = 7800\ \mathrm{cm}^2/\mathrm{mol}$.

vii. **Ninhydrin test according to Kaiser—detection of primary amines**

The ninhydrin test according to Kaiser is used to check the Fmoc decoupling for completeness:

Reagents:

- Solution 1: KCN in pyridine (2 mL 0.001 M KCN dissolved in 98 mL pyridine)
- Solution 2: 5% ninhydrin in butanol (w/v)
- Solution 3: 80% phenol in butanol (w/v)

Some dry resin beads are mixed with two drops each of solution 1, 2, and 3 in a reaction vessel. The reaction vessel is then heated to 110 °C for 5 min in an oven.

The coloration of resin and solution suggests the decoupling yield:

Colouring of the resin beads	Colouring of the solution	% responds
Light blue	Traces of blue	99.4%
Blue	Light blue	94.0%
Dark blue	Blue	84.0%
Dark blue	Dark blue	76.0%

viii. **Chloranil test—detection of secondary amines**

To verify complete Boc protection of a secondary amine, the choranil test is used:

- Solution 1: 2% *para*-chloroanil in DMF (w/v)
- Solution 2: 2% acetaldehyde in DMF (v/v)

A few dry resin beads are mixed with one drop each of solutions 1 and 2 in a reaction vessel. The reaction vessel is then incubated for 10 min at RT.

A blue coloration of the resin indicates the presence of secondary amines.

10.4 Protein Synthesis

Stepwise synthesis on the solid phase has proven successful for small peptides with 2 to 70 amino acid residues. Larger peptides are nevertheless difficult to synthesize with this method since globular structures form with increasing peptide length, which can prevent the complete acylation of the *N*-terminus. As a result, peptide mixtures of different lengths are formed (sometimes only a difference of one to two amino acids), which are difficult to separate due to their small chemical differences.

Various methods have been established for the synthesis of longer peptides. The largest group is called fragment condensation. This means that two shorter peptides are synthesized using solid-phase chemistry and joined together in a second separate step via a condensation reaction. These include native chemical ligation (NCL), Staudinger ligation, intein ligation, and sortase ligation. These different condensation reactions are often used orthogonally to each other in order to synthesize polypeptides with more than 300 amino acids.

Alternatively, longer polypeptides are produced using recombinant protein expression in bacteria or eukaryotic cells and subsequently purified. However, post-translational modifications cannot be introduced specifically enough with this method. Therefore, for specifically modified proteins, peptide synthesis, fragment condensation, and recombinant protein expression are usually the only options.

10.4.1 Native Chemical Ligation

The best-known fragment condensation in peptide chemistry is native chemical ligation (NCL). It was described as early as 1953 but has only been widely used since the 1990s and with the establishment of solid-phase peptide chemistry.

Specifically, an unprotected peptide fragment whose *C*-terminus is activated by a thioester selectively reacts with another peptide fragment bearing a cysteine at the *N*-terminus to form a new peptide bond with each other (Fig. 10.23).

The *C*-terminal thioester is not sufficiently activated that it can react with a nucleophilic amino group (*N*-terminus of another peptide fragment or a side-chain lysine) to form an amide or with a hydroxyl group to form an ester. However, the thioester can transesterify with another thiol. In the case of transesterification with the *N*-terminal cysteine, there is an entropically favored intramolecular reaction of the free amino acid with the thioester to form a peptide bond. This is referred to as the *S* → *N*-acyl rearrangement. The reaction proceeds in an aqueous solution. Usually, aromatic thiols such as thiophenol are added as catalysts to prevent the oxidation of the cysteine to cystine (formation of disulfides) and the incipient hydrolysis of the thioester in the aqueous solution.

In general, all amino acids can be used as *C*-terminal thioesters. However, valine, isoleucine, and proline show significantly slower conversion rates, as they are sterically

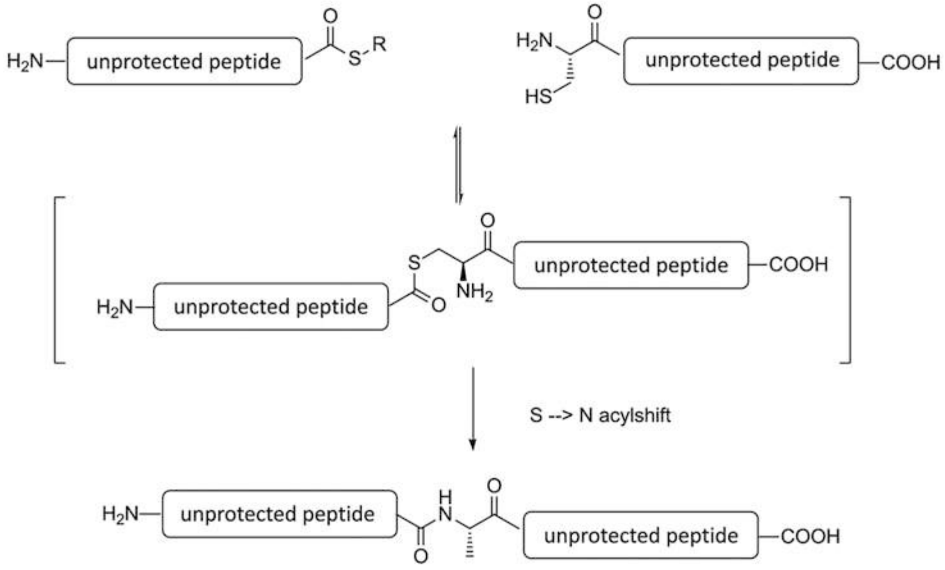


Fig. 10.23 Reaction course of native chemical ligation

hindered. Since thioesters are base labile, the synthesis in the Fmoc-SPPS is carried out via the Kenner linker.

10.4.2 Staudinger Ligation

Native chemical ligation allows the total chemical synthesis of long peptides. However, the method is limited: One needs a cysteine residue at the ligation site.

Another peptide fragment condensation is based on the Staudinger reaction. This does not require a cysteine residue and can therefore be used in a complementary manner to the native chemical ligation.

In the Staudinger reaction, an azide is reduced to an amine by the oxidation of triphenylphosphine. This is exploited in the peptide fragment condensation by introducing a thiol group in the *ortho-position in* one of the phenyl rings of the triphenylphosphine. Analogous to the native chemical ligation, the *C*-terminus of peptide A to be coupled is again present as a thioester. Here, as in the native chemical ligation, a transesterification with the thiol group of the triphenylphosphine takes place. In the next step, peptide fragment B is added, whose *N*-terminal amino group has been activated to form the azide (-N_3). The azide reacts with the phosphine to form an iminophosphorane with cleavage of an N_2 molecule. The nitrogen atom of the iminophosphorane reacts entropically favored intramolecularly with the thioester to form a peptide bond to form an amidophosphinium

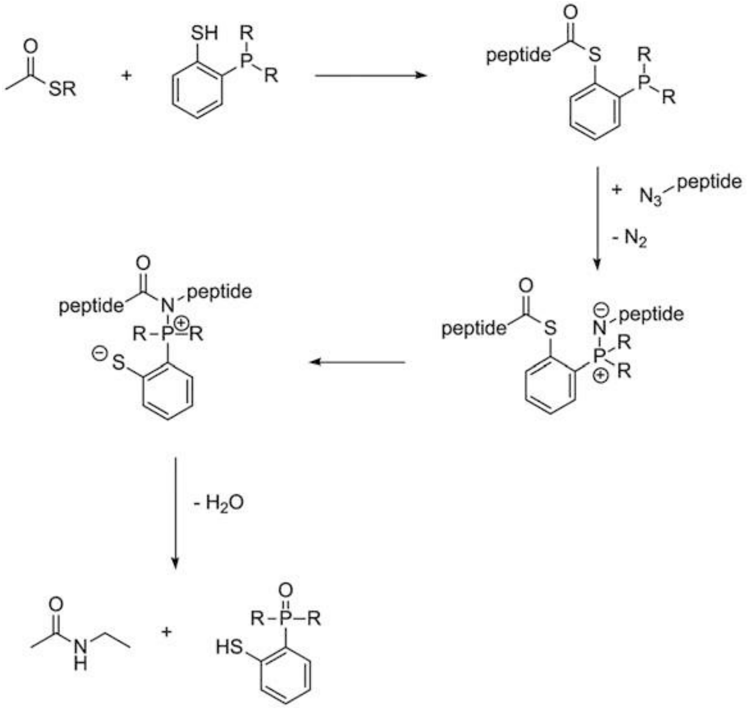


Fig. 10.24 Reaction course of Staudinger ligation

salt. In the final step, the amidophosphinium salt hydrolyzes, cleaving off the peptide formed and oxidizing the phosphorus to the phosphine oxide (Fig. 10.24). No atoms of triphenylphosphinethiol remain in the peptide product.

It should be noted that Staudinger ligation enables the total synthesis of long peptides orthogonally to native chemical ligation (Fig. 10.25). Therefore, these two techniques are usually used together. Native chemical ligation is linked to coupling via an *N*-terminal cysteine amino acid. While the chemically more complex Staudinger ligation (synthesis of a peptide thioester and an azido peptide) is possible independently of a cysteine residue.

10.4.3 Inteine

An intein is a section of a protein that can cut itself out and join the remaining parts (the exteins) with a peptide bond. This process is called protein splicing in analogy to RNA splicing. Inteins are, therefore "protein introns".

The mechanism of intein-mediated protein splicing is similar to that of native chemical ligation (Fig.10.26). The precursor protein contains three segments: an *N*-extein, an intein, and a *C*-extein. At the *C-terminus of* the *N-extein*, an *N* → *O* or *N* → *S shift* occurs when the

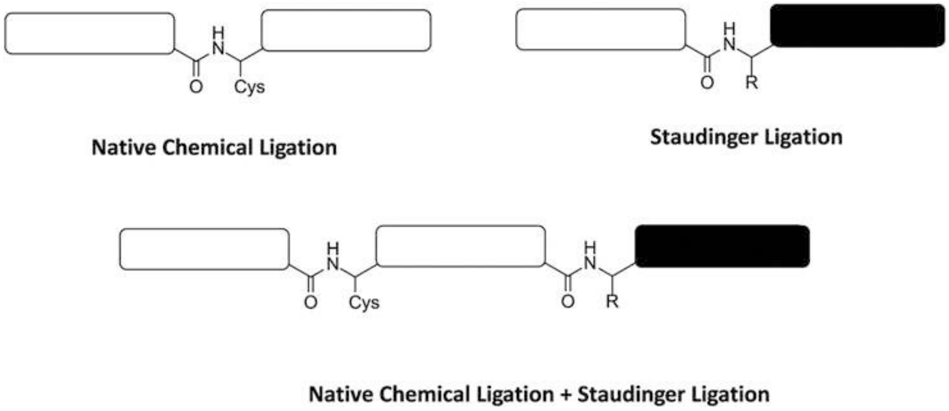


Fig. 10.25 Native chemical ligation and Staudinger ligation are orthogonal to each other. While coupling via NCL is dependent on a *C*-terminal cysteine, Staudinger ligation can occur independently of a second cysteine in the peptide

side chain of the *N*-terminal amino acid of the intein (a serine, threonine, or cysteine) nucleophilically attacks the carbonyl group of the peptide bond. The resulting ester is also nucleophilically attacked by the *C*-terminal residue (a serine, threonine, or cysteine). The result is transesterification. In the next step, the nitrogen atom of the *C*-terminal asparagine side chain of the intein attacks the carbonyl group of the peptide bond between the intein and the *C*-extein. In this process, the asparagine residue cyclizes to form a succinimide. In this process, the *N*- and *C*-extein, which are esterified with each other, are cleaved off. In a final step, an *O/S* → *N*-acyl shift occurs. A peptide bond is formed between the *N*-and *C*-extein.

Inteins have a length of 138 to 844 amino acids. Therefore, they are not synthesized by means of the Fmoc-SPPS but expressed in cells. In particular, toxic proteins can be produced in cells in this way, which are only later processed into their toxic form. Some intein reactions can be initiated by thiols or by lowering the pH.

10.4.4 Sortase-Mediated Ligation

Sortase is originally a prokaryotic enzyme that selectively links proteins to other proteins on the bacterial surface. Sortases occur mainly in Gram-positive bacteria. The transpeptidase activity of sortase is used to produce fusion proteins. The recognition motif (LPXTG) is attached to the *C*-terminus of protein A, while an oligo-glycine motif is attached to the N-terminus of protein B. Upon addition of the sortase to the two proteins, the two peptides are covalently linked via a native peptide bond.

The sortase recognizes the *C-terminal* peptide sequence LPXTG-XX (Leu-Pro-X-Thr-Gly) of the target protein. The sortase then cuts between threonine and glycine and forms a

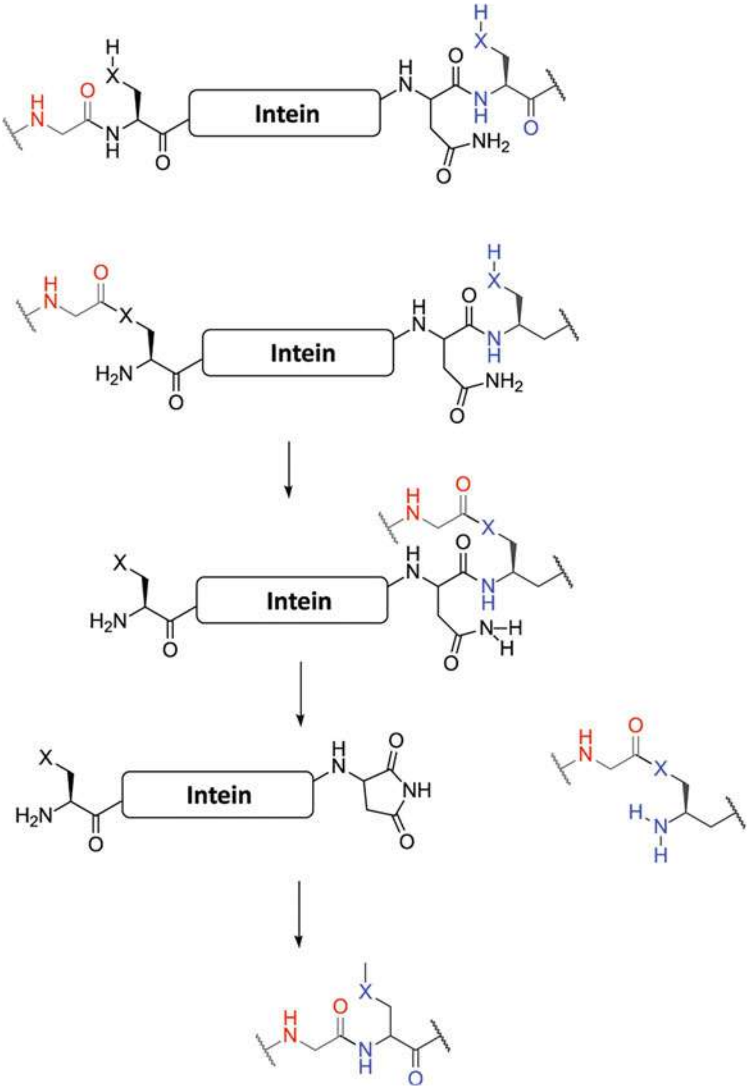


Fig. 10.26 Reaction course of an intein ligation

thioester with the *C*-terminus of threonine. In the next step, the sortase attaches with the threonine thioester to a highly hydrophobic transmembrane sequence consisting of a sequence of several glycine residues. The *N*-terminus of the $(G)_n$-peptide nucleophilically attacks the thioester, resulting in the fusion peptide with the middle sequence segment LPXT-$(G)_n$ (Fig. 10.27).

As a disadvantage, however, it should be mentioned that sortase-mediated ligation is coupled to the LPXT-$(G)_n$ peptide sequence. If a fusion protein with a determined sequence is required, sortase-mediated ligation is therefore useless.

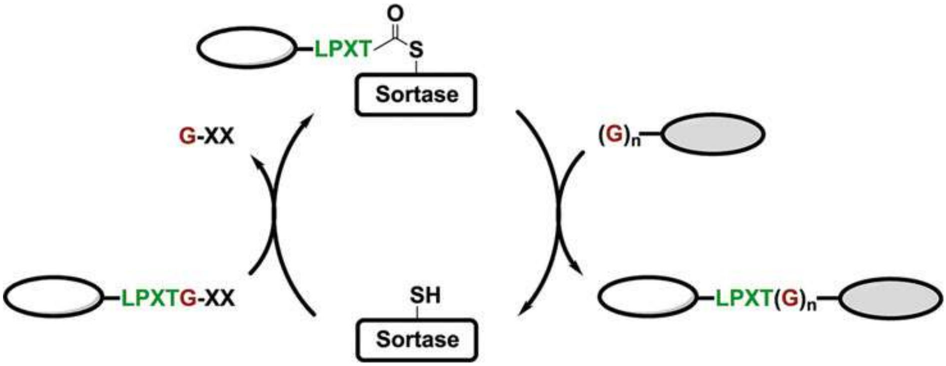


Fig. 10.27 Sortase-mediated ligation is analogous to NCL via an activated thioester. However, one is bound to the sequence LPXT(G)$_n$

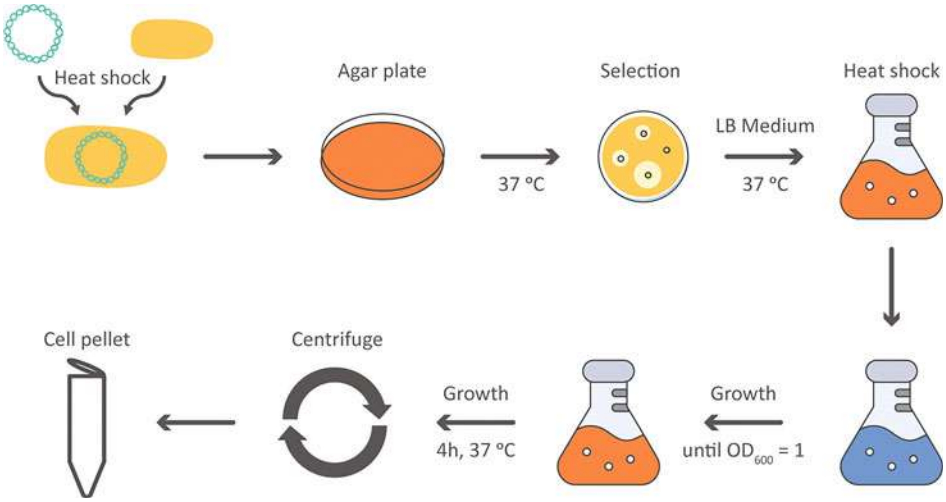


Fig. 10.28 Schematic representation of protein expression in *E. coli*

10.4.5 Recombinant Proteins

Currently, the simplest method to synthesize a protein with more than 100 amino acids is recombinant protein expression. Usually, the production of the target protein is achieved by manipulating gene expression in an organism so that this organism only produces the target protein.

Several gene expression systems exist. However, recombinant gene expression dominates in *E. coli.* As shown in Fig. 10.28, the DNA of the protein to be synthesized in the form of a plasmid (ring-shaped DNA) is incubated with the bacteria under heat. This allows the plasmid to enter the interior of the bacteria. The plasmid encodes the target protein, the lac promoter, and an antibiotic resistance gene. The latter enables the bacteria to grow on agar that is provided with an antibiotic. The reason for this is that only bacteria that have taken up the plasmid can grow. The grown bacterial colonies are then picked and suspended in LB medium (*lysognybroth*). When the bacterial growth reaches a certain density, the addition of the lactose analog isopropyl-ß-D-thiogalactopyranoside (IPTG) activates the lac promoter. The bacteria increasingly produce only the encoded target protein. After approximately 4 h of growth, the bacterial solution is centrifuged. The pellet obtained is digested by cell lysis or pressing and the protein obtained is purified by chromatographic methods.

Recombinant protein expression allows the production of long polypeptides in large quantities. Gene expression in prokaryotic systems such as *E. coli* offers no possibility at all of the post-translational modifications to the side chain of the protein. And even the use of eukaryotic systems such as Chinese hamster ovary (CHO) cells allows only limited control over which residues are post-translationally modified or not. Therefore, despite advances in recombinant protein expression, the total synthesis of post-translationally modified proteins using peptide solid-phase chemistry and the ligation methods mentioned above is of high interest.

10.5 Post-Translational Modifications

Post-translational modification (PTM) refers to the covalent, usually enzymatically mediated, modification of proteins after protein biosynthesis.

In protein biosynthesis, the sequence information of the mRNA is translated into a polypeptide chain. Subsequently, this polypeptide chain can be further modified. The post-translational modification increases the chemical diversity of the protein and may therefore be necessary for folding and, consequently, for a new function of the protein.

Post-translational modifications are divided into two groups: Covalent modifications of amino acid residues and cleavage of the protein backbone (Fig. 10.29).

Post-translational modifications occur at reactive amino acid side chains or the *C* or *N* termini of the protein. As shown in Fig. 10.30, the five most common covalent

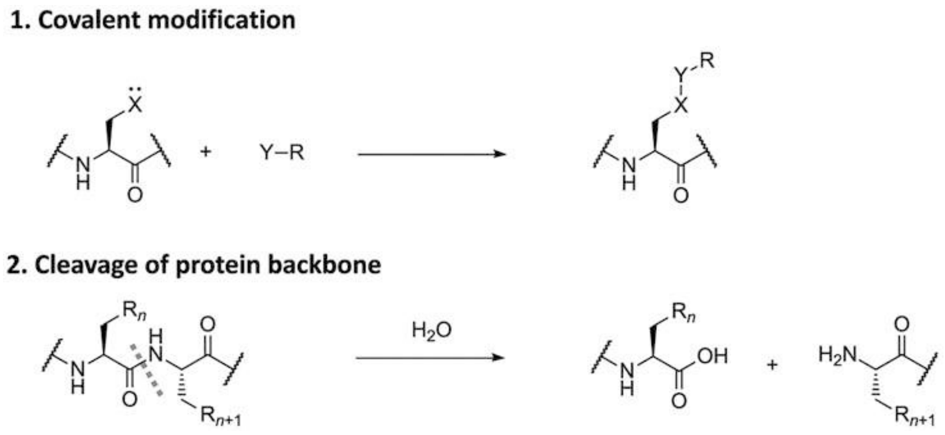
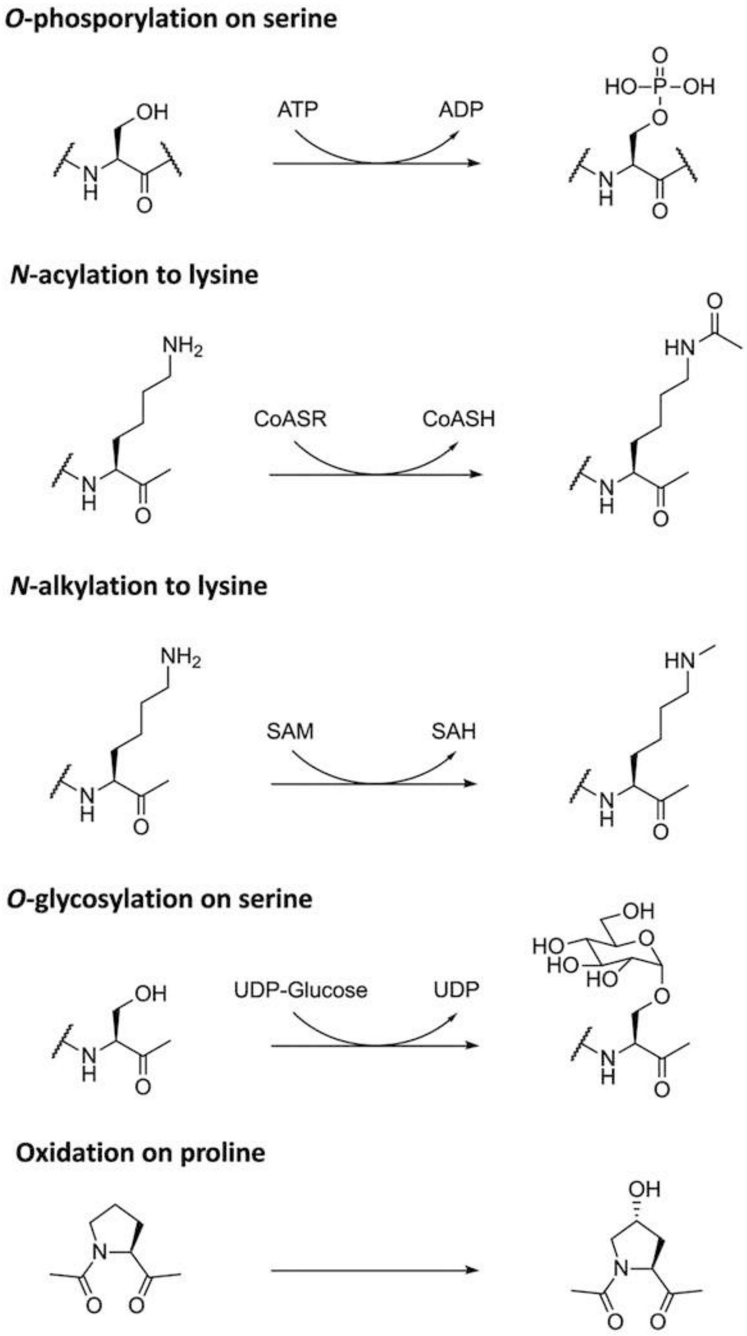


Fig. 10.29 Classification of post-translational modification into the covalent modification of the side chain and irreversible cleavage of the protein backbone



O-phosporylation on serine

N-acylation to lysine

N-alkylation to lysine

O-glycosylation on serine

Oxidation on proline

Fig. 10.30 The five main groups of post-translational modifications: *O*-phospholation, *N*-acylation, *N*-alkylation, *O*-glycosylation, and oxidation of the proline ring

modifications of side-chain residues are 1) phosphorylation of the hydroxyl groups on serine, threonine, or tyrosine side chains; 2) acylation of the lysine side chain or *N*-terminus; 3) alkylation of the lysine side chain or *N*-terminus; 4) glycosylation of the serine or 5) the oxidation of the proline ring by the introduction of a hydroxy group.

Phosphorylation of side-chain residues is particularly useful for protein regulation. Nevertheless, kinases (phosphorylating enzymes) and phosphatases (dephosphorylating enzymes) are drug targets in different indications.

It should be noted that the covalent post-translational modification is reversible in most cases by enzyme catalysis. In contrast, cleavage of the peptide backbone is irreversible. In these cases, an inactive precursor protein, so-called zymogen or proenzyme, is cleaved into the active enzyme. A well-known example is the initiation of apoptosis, the regulated cell death, by the cleavage of procaspase into the active enzyme caspase.

10.6 Summary

For a long time, peptides and proteins could only be isolated from animal tissue. Nowadays, peptides up to 20 amino acids in length can be rapidly synthesized in the solid phase. Peptides up to 100 amino acids in length are accessible by fragment condensation. Longer peptides and proteins are produced using protein expression. In contrast, post-translationally modified proteins are possible using gene expression in eukaryotic systems but selective modification is still difficult. Specifically, peptides as probes and non-post-translationally modified proteins can be readily produced and are available for chemical biology experiments.

Proteins as Drug Targets 11

Investigations of the approved drugs for their drug targets have shown that they mostly bind to proteins and, in particular, to receptors or enzymes (Fig. 11.1). There is a mechanistic and a structural reason for this. The mechanistic reason is that proteins are the function carriers in the organism. Manipulation of their function has direct, that is, clinical, effects. In this context, receptors and enzymes form the largest group of drug targets due to their structure. Receptors and enzymes form binding pockets for their native ligands. This is exploited in drug discovery. Many chemical probes are based on natural ligands and substrates and are therefore chemically structurally very similar to them.

11.1 Receptors

Receptors are proteins that recognize endogenous signals in the form of a chemical messenger and thus transmit, amplify or integrate the signal (Fig. 11.2). Transduction of the signal amplifies the effect of the single ligand. Integration can initiate the signal into another biochemical pathway.

The receptor proteins are distinguished based on their localization as follows:

Ion Channel-Bound (Ionotropic) Receptors

Ion channel-bound receptors are targets of fast neurotransmitters, such as acetylcholine or gamma-aminobutyric acid (GABA). Activation of these receptors leads to changes in ion movement across a membrane. These receptors have a heteromeric structure in that each subunit consists of the extracellular ligand-binding domain and a transmembrane domain, with the transmembrane domain in turn containing four transmembrane alpha-helices.

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_11

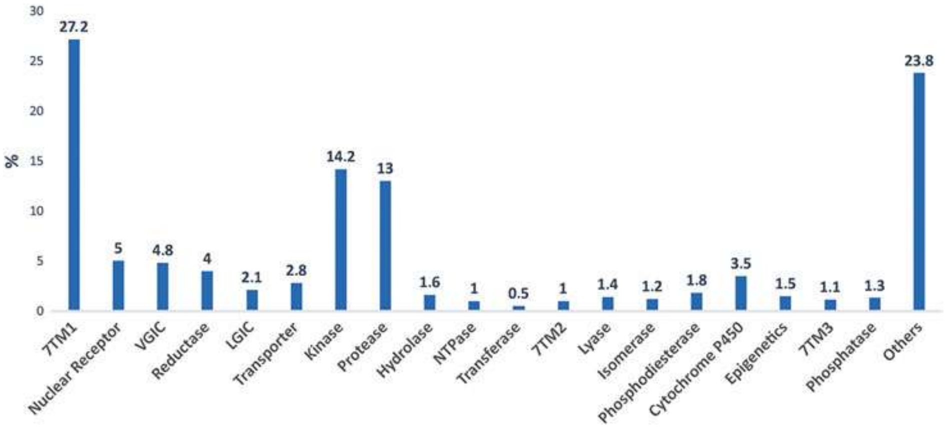


Fig. 11.1 Percentage distribution of protein drug targets among their subclasses. Protein drug targets are mostly receptors (27.2%), enzymes such as kinases (14.2%) or proteases (13%), and others (23.8%)

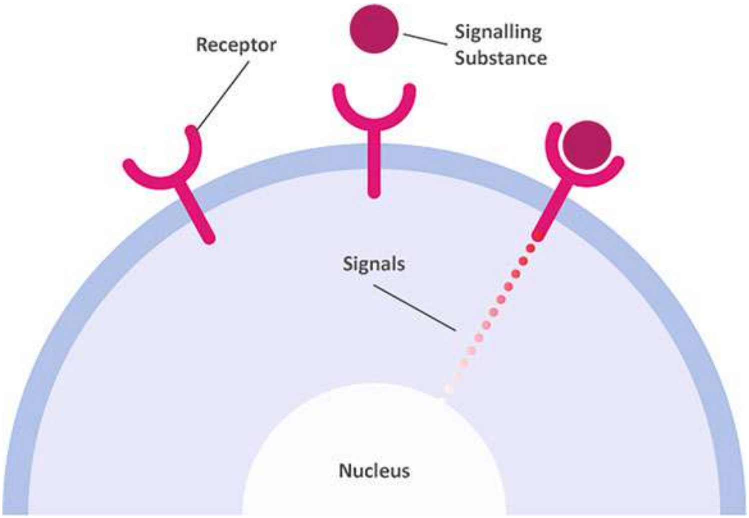


Fig. 11.2 Principle of signal transmission at a cell surface receptor

Hormone-Linked (Metabotropic) G-Protein Coupled Receptors (G-Coupled Receptor: GPCR)

The G protein-coupled receptors form the largest family of receptors. Hormones, such as dopamine or glutamate bind to them. They consist of seven transmembrane

(continued)

alpha-helices. The alpha helices connecting loops form extracellular and intracellular domains. The binding site for larger peptide ligands is usually in the extracellular domain, whereas the binding site for smaller non-peptide ligands is often between the seven alpha-helices and an extracellular loop. As the name implies, the receptors are coupled to various intracellular effector systems via G-proteins for signal transduction.

Enzyme-Linked Receptors

Enzyme-linked receptors (mostly kinase-linked receptors) consist of an extracellular domain containing the ligand-binding site and an intracellular domain, often with enzymatic function, linked by a single transmembrane. The best-known example is the insulin receptor.

Core Receptors

Although they are called nuclear receptors, they are located in the cytoplasm and migrate to the nucleus only after binding with their ligands. They consist of a *C*-terminal ligand-binding region, a DNA binding domain, and an *N*-terminal domain. The DNA binding domain has two zinc fingers that are responsible for recognizing DNA sequences specific to that receptor. The *N*-terminus interacts with other cellular transcription factors in a ligand-independent manner. The best-known examples are the steroid and thyroid hormone receptors.

Ligand-Receptor Binding

The binding of a ligand to a protein receptor is usually non-covalent and therefore reversible. The interactions between the ligand and the receptor are based on

- hydrogen bonds,
- electrostatic,
- Van der Waals interactions and
- hydrophobic effects.

Ligand binding occurs according to the law of mass action in an equilibrium process of unbound and bound states. For the concentrations of the receptor [R], the ligand [L], and the receptor-ligand complex [RL] the following relationship results:

$$[R] + [L] \rightarrow [RL]$$

$$K_D = \frac{[R][L]}{[RL]}$$

The resulting dissociation constant K_D describes the binding property or affinity of the ligand L to the receptor R. This corresponds to the ligand concentration [L] in molar (mol/L or M) when half of the protein concentration has bound the ligand. The smaller the dissociation constant (in the nano- or picomolar range), the better the ligand binds to the receptor.

In pharmacology, however, a distinction must be made between the affinity and the efficacy of a ligand:

Affinity is the ability of a drug to bind to a receptor to form the ligand-receptor complex.

Efficacy is the ability of a drug-receptor complex to elicit a response.

To experimentally determine the dissociation constant K_D of a ligand, the measured occupancy of the receptor is plotted against the concentration of the ligand. A characteristic saturation curve is obtained (Fig. 11.3). The curve intersects at the level of 50% occupancy at the concentration of the ligand corresponding to the dissociation constant. If the concentration of the ligand is plotted in logarithmic steps, a sigmoid function is obtained whose inflection point corresponds to the concentration of the dissociation constant K_D. The latter method was the standard before computerized non-linear regression software became available.

Ligands, Agonists, Allosteric Modulator, and Antagonists

Not every ligand that binds to a receptor also activates this receptor and leads to a pharmacological effect. The following classes of ligands exist (Fig. 11.4):

Agonists

Agonists activate the receptor. The natural ligand with the greatest efficacy for a receptor is an agonist (100% efficacy).

Partial agonists do not activate the receptor with maximal efficacy, even with maximal binding, resulting in partial responses compared to those of full agonists (efficacy between 0 and 100%).

Allosteric Modulators

Allosteric modulators do not bind to the agonist binding site of the receptor but to specific allosteric binding sites via which they modify the effect of the agonist. A well-known example is benzodiazepines, which bind to an allosteric binding site of the GABA receptor and thus enhance the effect of natural GABA.

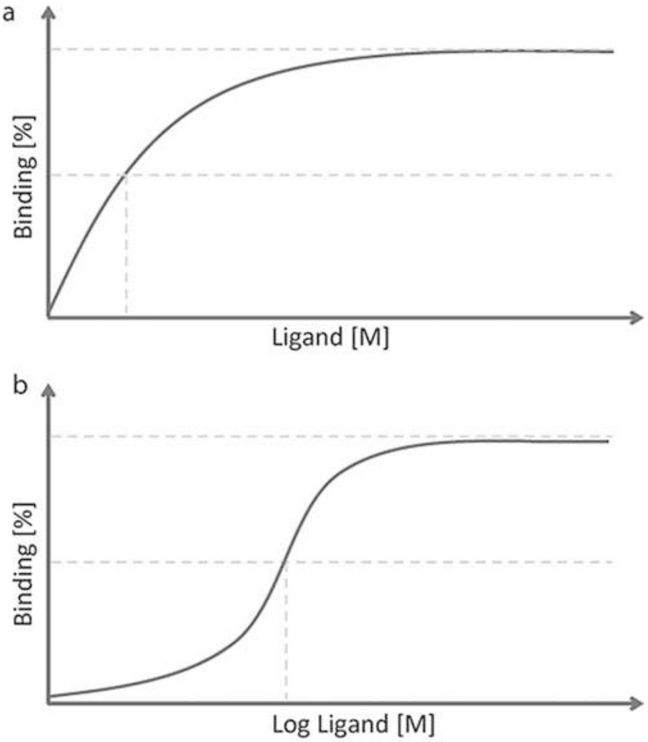


Fig. 11.3 Plot of the bound ligand concentration: **a** non-logarithmic and **b** logarithmic. If the ligand concentration is plotted logarithmically, a saturation curve is obtained. If, on the other hand, the ligand concentration is plotted logarithmically, a sigmoid curve is obtained. Reading the dissociation constant K_D at the inflection point of the sigmoid curve is much easier than with the saturation curve

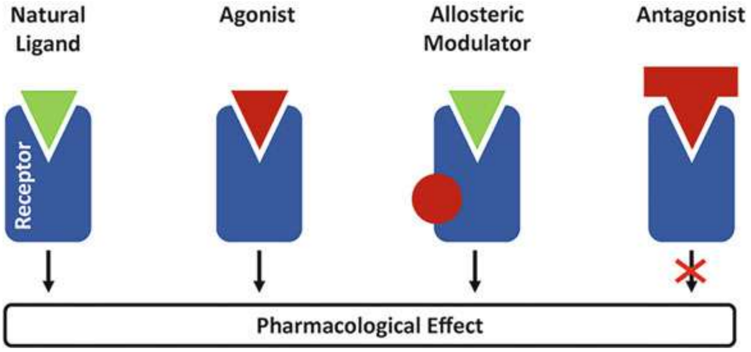


Fig. 11.4 Overview of natural ligands, agonists, allosteric modulators, and antagonists and their pharmacological effect

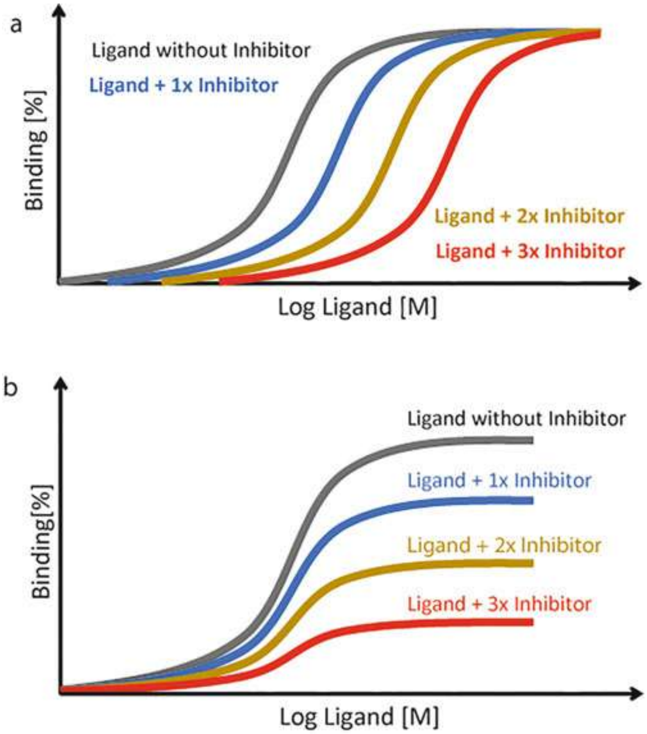


Fig. 11.5 (**a**) Inhibition binding curves of a competitiveantagonist, (**b**) Inhibition binding curves of an uncompetitive, irreversibleantagonist

Antagonists

Antagonists bind to receptors but do not activate them. This leads to receptor blockade, which inhibits the binding of agonists and inverse agonists. Receptor antagonists can be competitive (reversible) and compete with the agonist for the receptor (Fig. 11.5a), or they can be irreversible antagonists that form covalent bonds (or noncovalent bonds with extremely high affinity) with the receptor and block it completely (Fig. 11.5b). These are then not in competition with the natural ligands, which is called uncompetitive inhibition. An example of an irreversible antagonist is the proton pump inhibitor omeprazole. The effect of the irreversibly bound omeprazole can only be reversed by the re-synthesis of the receptor.

Inverse Agonists

An inverse agonist binds to the same receptor as an agonist but elicits the opposite pharmacological response as the agonist. In relation to the agonist, one then speaks of a negative effect of the inverse agonist.

However, the term is sometimes used in pharmacology to refer to other proteins that are drug targets, such as enzymes, transporters, and ion channels.

11.2 Enzymes

$$\mathrm{A} + \mathrm{B} \underset{\mathrm{k}_{-1}}{\overset{\mathrm{k}_1}{\rightleftharpoons}} \mathrm{C} + \mathrm{D} \tag{11.1}$$

Enzymes are proteins that catalyze chemical reactions. They influence the rate at which a reaction reaches equilibrium by lowering the activation energy (Fig. 11.6). This becomes clear when one looks at the Arrhenius equation, i.e. the relationship between temperature and reaction rate:

$$k = \mathrm{Ae}^{-\mathrm{Ea/RT}} \tag{11.2}$$

In Eq. 11.2, the term A is the maximum rate constant that the reaction would have if all the molecules possessed sufficient collision energy to apply the activation energy. At very high temperatures, the exponent is small, $\mathrm{e}^{\text{-Ea/RT}}$ approaches 1, and *k,* the reaction rate, becomes almost equal to A.

Accordingly, enzymes cannot shift the equilibrium position of a reaction, defined by the equilibrium constant K', i.e., they can only increase the rate constants for both the outward and reverse reactions to the same extent.

In equilibrium, therefore:

$$V_{\mathrm{forth}} = V_{\mathrm{back}} \ (11.3)$$

$$\text{Towards reaction}: \quad V_{\mathrm{forth}} = k_1 \cdot [A] \cdot [B] \tag{11.3a}$$

$$\text{Back reaction}: \quad V_{\mathrm{back}} = k_{-1} \cdot [C] \cdot [D] \tag{11.3b}$$

$$\frac{k_1}{k_{-1}} = \frac{[C] \cdot [D]}{[A] \cdot [B]} = K' \ (11.3\mathrm{c})$$

Nevertheless, the question arises as to how enzymes are able to lower the activation energy and thus accelerate the back and forth reaction. The answer to this question lies in the binding energy that results from the formation of multiple weak bonds and interactions between the enzyme and its substrate. The formation of this enzyme-substrate complex is first mediated by weak binding energies that compensate for the loss of entropy of the substrate. In the next step, the interactions formed are optimized in the transition state of the reaction. That is, the active site is not generally complementary as earlier theories of the lock-and-key principle state but it is complementary to the transition state that the substrate undergoes as the reaction proceeds to the product. This is because binding energy is the main source of free enthalpy that enzymes can draw upon to lower the activation energy of a reaction. In addition, the enzyme itself is thought to undergo a conformational change. In

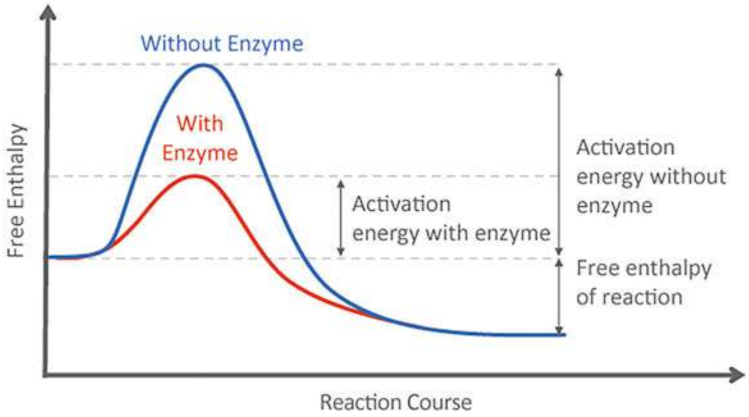


Fig. 11.6 Free enthalpy reaction diagram for a non-catalyzed (blue line) and an enzyme-catalyzed reaction (red line). The deeper the cut in the energy profile of the enzyme-catalyzed reaction, the more stable is the intermediate in question

this way, specific functional groups on the enzyme can be brought into an optimal position for catalysis. This process is referred to as *induced fit*.

11.2.1 Enzyme Classification

Enzymes are classified into six enzyme classes (EC: *enzyme classification*) according to the biochemical reactions they catalyze. Each enzyme is assigned four numbers, the order of which is defined internationally (Table 11.1).

Examples

EC 1: Oxidoreductases

Catalysis of reduction and oxidation reactions

Eg: Lactate dehydrogenase (lactate: NAD$^+$ oxidoreductase)

Donor Acceptor EC

EC 2: Transferases

Catalysis of the transfer of atomic groups

Eg: Hexokinase (ATP: glucose-6-phosphotransferase)

Donor Acceptor EC

EC 3: Hydrolases

Catalysis of the hydrolysis of the substrate

Eg: Glucose-6-phosphatase (Glucose-6-phosphate hydrolase)

(continued)

Table 11.1 The six enzyme classifications and the nature of their catalytic function

Enzyme class	Type of catalytic reaction
1. Oxidoreductases	Oxidation-reduction reactions
2. Transferases	Transfer of functional groups
3. Hydrolases	Hydrolase reactions
4. Lyases	Elimination of groups with formation of double bonds
5. Isomerases	Isomerizations
6. Ligases	Bond formation coupled with ATP hydrolysis

EC 4: Lyases

Catalysis of the non-hydrolytic cleavage of atomic groups from a molecule

Eg: Glutamate decarboxylase

EC 5: Isomerases

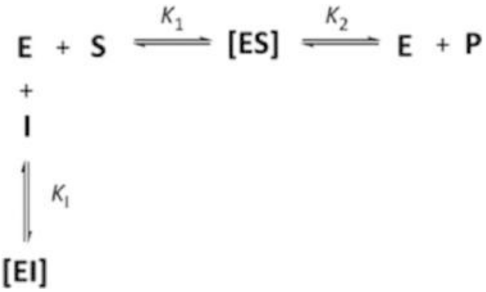
Catalysis of changing the structure of the substrate without changing its composition:

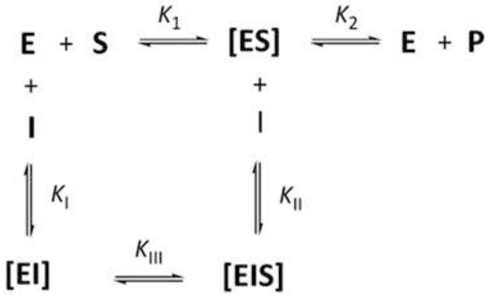
(a) Mutases intramolecular group changes
(b) Epimerases epimerization of an optically pure compound
(c) Racemases racemization of an optically pure compound

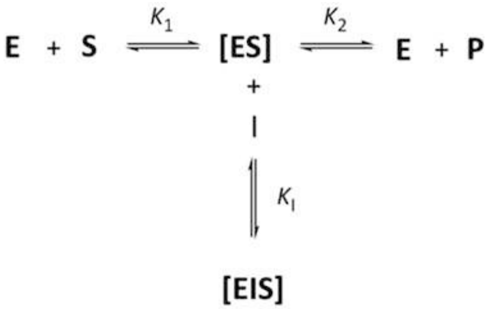
EC 6: Ligases.

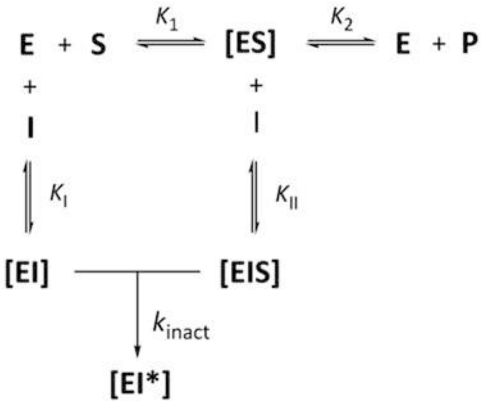
Catalysis of a bond linkage of two molecules under the cleavage of a high-energy compound:

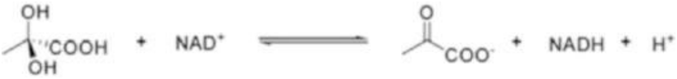
Eg: Acetyl-CoA synthetase




E + S
K_1
[ES]
K_2
E + P
+
I
+
I
K_I
K_{II}
[EI]
K_{III}
[EIS]


E + S
K_1
[ES]
K_2
E + P
+
I
K_I
[EIS]


E + S
K_1
[ES]
K_2
E + P
+
I
+
I
K_I
K_{II}
[EI]
[EIS]
k_{inact}
[EI*]


OH
COOH
OH
+
NAD⁺
O
COO⁻
+
NADH
+
H⁺

11.2.2 Special Catalysis Mechanisms of Enzymes

In addition to the concept of enzyme catalysis already introduced in Sect. 11.2.1, this is generally supported by suitably positioned catalytically active groups, depending on the substrate and the reaction to be catalyzed. The catalytically active groups support catalysis by the following mechanisms:

- general acid-base catalysis,
- covalent catalysis, and
- metal-ion catalysis.

None of the three mechanisms occurs independently of the other. On the contrary, the interaction of these mechanisms is indispensable for catalysis.

Acid-Base Catalysis

In general acid catalysis, a partial proton transfer from an acid lowers the free enthalpy of the transition state of a reaction.

For example, an uncatalyzed keto-enol tautomerization (Fig.11.7) is on the side of the more stable ketone because of the high energy (free enthalpy) of the carbanionoid transition state. However, protonation of the oxygen atom reduces the carbanion character of the transition state, thereby accelerating the reaction to the enol.

A reaction may also be stimulated by general base catalysis if its rate is increased by partial deprotonation. Some reactions can undergo both processes simultaneously, these are concerted acid-base catalyzed reactions.

The characteristic of acid-base catalysis in enzymes is above all the amino acid side chain of histidine (Fig. 11.8; catalytic triad). The side chain of histidine has a pK value of 6.0. Histidine, therefore, serves as an ideal acid-base catalyst at physiological pH.

Covalent Catalysis

The core of covalent catalysis is the generation of a short-lived covalent bond between enzyme and substrate. An example of this is the hydrolysis of the bond between A and B.

$$A - B \quad \rightarrow H_2OA + B$$

In the presence of a nucleophilic function X of the enzyme, the covalent transition structure between A and the enzyme X is formed:

$$A - B + \ :X \rightarrow A - X + B \rightarrow H_2OA + \ :X + B$$

The resulting compound A − X is significantly more reactive for the subsequent reaction than compound A − B. The basis of this form of catalysis is the faster sequence of both reactions compared to the uncatalyzed hydrolysis of A − B. This catalysis is usually mediated by amino acids with nucleophilic residues. Nevertheless, it is decisive that the free enzyme must always be regenerated by a further reaction.

An example in which general acid-base catalysis and covalent catalysis occur together is the catalytic triad of serine proteases. Figure 11.8 shows the catalytic mechanism. The reaction involves the nucleophilic attack of the serine in the active site on the C atom of the

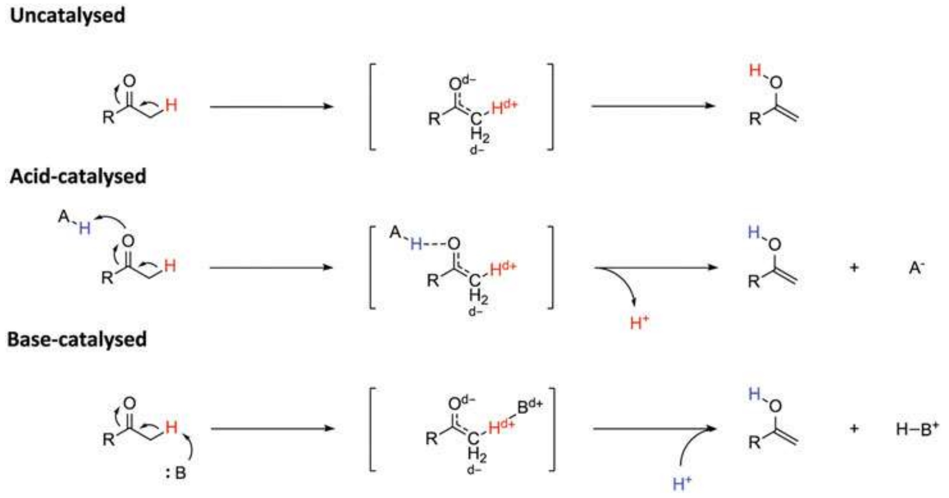


Fig. 11.7 Mechanism of keto-enol tautomerism. (**a**) not catalyzed, (**b**) general acid catalysis, (**c**) general base catalysis. H-A stands for acid; B for base

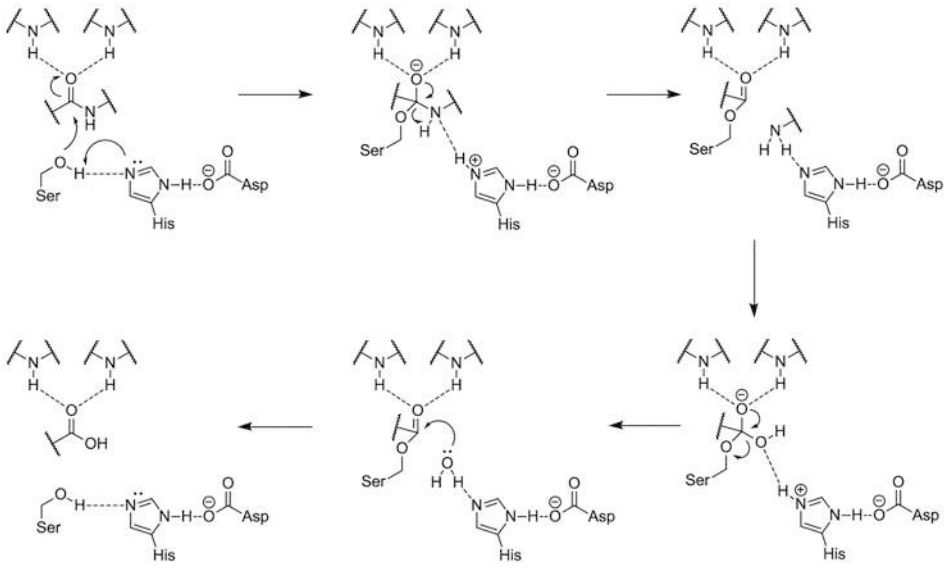


Fig. 11.8 The catalytic mechanism of serine proteases

bond to be cleaved to form the tetrahedral transition state, the conversion of the tetrahedral transition state to the acyl-enzyme intermediate by general acid catalysis by the histidine polarized by the aspartate in the active site. Followed by the departure of the amino product and its replacement by a water molecule, the reversal of step 2 leading to a second tetrahedral intermediate, and the reversal of step 1 to release the carboxyl product of the reaction and recover the active enzyme.

Metal Ion Catalysis

In addition to the other two catalytic mechanisms, metal ions play a particularly important role. About one-third of all enzymes require the presence of metal ions. A distinction is made between metalloenzymes, which, for example, catalyze the soft Lewis acids.

Fe^{2+}, Fe^{3+}, Cu^{2+}, Zn^{+}, Mn^{2+}or Co^{2+},as cofactors, and on the other hand metal-activated enzymes, which primarily bind the hard Lewis acids.

Na^{+}, K^{+}, Mg^{2+}or $Ca^{2+}$for structural reasons.

In general, the catalytic mechanism of metal ions can be evaluated in different ways, so we distinguish three forms of catalysis of metal ions:

1. Binding to substrates to bring them into a conformation suitable for the reaction or to activate them as Lewis acids.
2. Reversible change in the oxidation state of metal ions to mediate redox reactions.
3. Stabilization of native protein conformation.

11.3 Enzyme Kinetics

The rate of enzyme-catalyzed reaction *V* varies with substrate concentration [*S*] for many enzymes, as Figs. 11.9 and 11.10 illustrate. Here, the rate *V* is defined as the number of moles of product formed per minute per volume. In the case of a constant enzyme concentration at low [*S*], *V* is linearly related to [*S*]; but if [*S*] is greater, *V* is effectively independent of [*S*].

Based on these observations, Leonor Michaelis and Maud Menten proposed a model in 1913 that explained this phenomenon mechanistically:

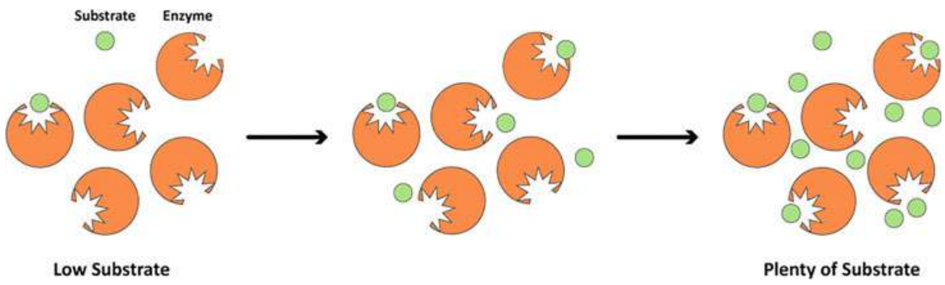


Fig. 11.9 Representation of the substrate binding of an enzyme as a function of the substrate concentration

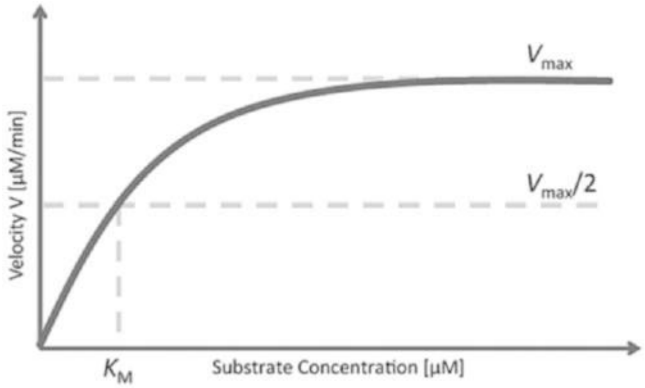


Fig. 11.10 Dependence of the catalysis rate on the substrate concentration

Overview

According to the model, a specific enzyme-substrate complex *ES* must occur as an intermediate during catalysis.

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

The enzyme *E* binds the substrate *S* to form an *ES complex* with a rate constant k_1. The enzyme-substrate complex can now, in turn, decompose into its starting products, $k - 1$, or convert the substrate *S* into the product *P*, k_2.

The initial velocity V_0 is determined by the degradation of the enzyme-substrate complex *ES* to the product. This process in turn depends on the concentration of the enzyme-substrate complex [*ES*]:

$$V_0 = k_2 \cdot [ES]$$
(11.4)

[*ES*] cannot be determined, therefore, the expression [E_t] is introduced for the total enzyme concentration (the sum of free and substrate-bound enzyme). The free enzyme then has the concentration [E_t]−[*ES*].

1. The formation and decay rates of *ES* are determined by steps for which the rate constants k_1(formation) and $k - 1 + k_2$(decay) are decisive.

$$V \text{ of the } ES - \text{Formation} = k_1 \cdot ([E_t] \cdot [ES]) \cdot [S]$$
(11.5)

$$V \text{ of the } ES - \text{Decay} = k_{-1} \cdot [ES] + k_2 \cdot [ES]$$
(11.6)

(continued)

2. The most important assumption for successfully describing the kinetics of an enzyme-catalyzed reaction is the *steady-state model.* The *steady-state model* means that the concentration of the intermediate product remains constant, while the concentrations of the starting materials and end products change (Fig. 11.11):

$$\frac{d[ES]}{dt} = 0$$

(11.7)

This assumes that the rates of formation and decay of *ES* are equal (see also Eqs. (11.5) and (11.6)):

$$V_{\mathrm{Formation}} = V_{\mathrm{Decay}}$$

(11.8)

$$k_1 \cdot ([E_{\mathrm{t}}] - [ES]) \cdot [S] = k_{-1} \cdot [ES] + k_2 \cdot [ES]$$

(11.8a)

3. Equation (11.8a) can now be solved for [*ES*]:

$$k_1 \cdot ([E_{\mathrm{t}}] - [ES]) \cdot [S] = k_{-1} \cdot [ES] + k_2 \cdot [ES]$$

(11.8a)

$$k_1 \cdot [E_{\mathrm{t}}] \cdot [S] - k_1 \cdot [ES] \cdot [S] = (k_{-1} + k_2) \cdot [ES]/ + k_1 \cdot [ES] \cdot [S]$$

(11.8b)

$$k_1 \cdot [E_{\mathrm{t}}] \cdot [S] = (k_1 \cdot [S] + k_{-1} + k_2) \cdot [ES]$$

(11.8c)

to [*ES*], the result is:

$$[ES] = \frac{k_1[E_t][S]}{k_1[S]+k_{-1}+k_2}$$

(11.9)

Summarizing the constants yields:

$$[ES] = \frac{[E_t][S]}{[S]+(k_2+k_{-1})/k_1}$$

(11.9a)

$$(k_2 + k_{-1})/k_1 = K_{\mathrm{M}}$$

(11.9b)

$$[ES] = \frac{[E_t][S]}{K_M+[S]}$$

(11.9c)

4. Substituting the term for [*ES*] from Eq. (11.9c) into Eq. (11.10), we get:

(continued)

$$V_0 = k_2 \cdot [ES]$$

(11.10)

$$V_0 = \frac{k_2[E_t][S]}{K_M+[S]}$$

(11.11)

5. At maximum conversion, i.e. maximum catalysis rate, this means that all enzymes are saturated with substrate. One can, therefore, state:

$$V_{max} = k_2 \cdot [\mathrm{E_t}]$$

Which, in turn, when used in Eq. (11.11), results in:

$$V_0 = \frac{v_{max}[S]}{K_M+[S]}$$

(11.12)

Equation (11.12) is also referred to as the MICHAELIS-MENTEN equation. This equation underlies the data of the kinetics shown in Fig. 11.10. At very low substrate concentrations, when [S] is significantly less than K_M, V = [S] $V_\mathrm{max}/K_\mathrm{M}$; which says nothing other than that the rate is directly proportional to the substrate concentration. At high substrate concentrations, when [S] is significantly greater than $\mathrm{K_{M,}}$ $V = V_\mathrm{max}$; i.e., the velocity is maximum, independent of the substrate concentration.

The meaning of K_M is already clear from Eq. (11.22). For [S] = K_M, V = $V_{max}/2$. Therefore, K_M corresponds to the substrate concentration at which the reaction rate reaches half its maximum value.

Pre-Steady-State Kinetics At the first moment, after an enzyme is mixed with the substrate, the enzyme-substrate complex is still forming. As shown in Fig. 11.11, d[ES]/dt $\neq$ 0.

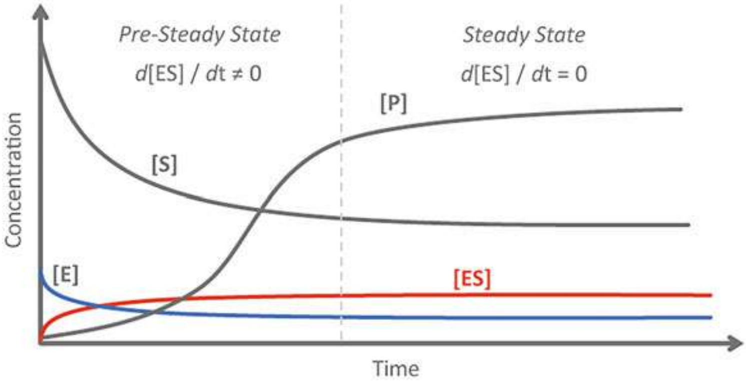


Fig. 11.11 Ratio of product and substrate concentration after time. Important here is the constant concentration of the enzyme-substrate complex *ES*

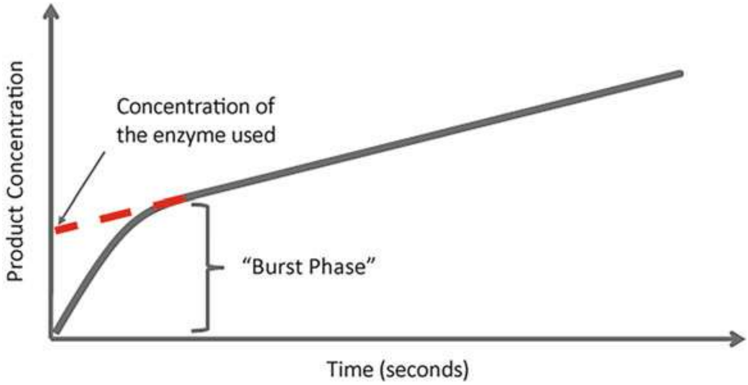


Fig. 11.12 *Pre-steady-state* curve showing the *burst phase* of an enzyme reaction

This state is called the *pre-steady state*. When the ES complex reaches equilibrium, i.e. d [*ES*]/dt = 0, it is called *steady-state*.

As shown in Fig. 11.12, the special feature of the *pre-steady-state kinetics*is the so-called burst phase that can be observed. Here, the product formation occurs very quickly and can be recognized by a compressed saturation curve at the beginning of the reaction. After the onset of the *steady-state,* the product formation turns into a straight line. In many fluorescence-based enzyme assays, such behavior is observable. However, it is not always immediately clear whether this is a burst phase or whether fluorescence quenching triggered by the substrate or an inhibitor is observed. In order to exclude fluorescence quenching, the straight line obtained is extended to the *y-axis* in order to read off the concentration of the enzyme used. If this agrees with the enzyme concentration used in the experimental prescription, fluorescence effects of the substrate or inhibitor can be excluded.

Lineweaver-Burk Diagram

The plot of *V* versus [S] in Fig.11.10 is not linear. Although it is initially linear at low [*S*], it bends upward to saturate at high [*S*]. Before the use of nonlinear curve-fitting programs on computers, the nonlinearity made it difficult to determine K_M and V_{max} accurately. Linearizations of the Michaelis-Menten equation were, therefore, developed. The best-known example is the Lineweaver-Burk diagram.

The Lineweaver-Burk diagram is obtained from the inverse of both sides of the Michaelis-Menten equation. As shown in Fig. 11.13, this is a linear form of the Michaelis-Menten equation and produces a straight line with the equation $y = mx + c$ with a *y*-axis intercept corresponding to $1/V_{max}$ and an *x*-axis intercept of the graph representing: $1/K_M$.

$$\frac{1}{v} = \frac{KM}{Vmax(S)} + \frac{1}{Vmax}$$

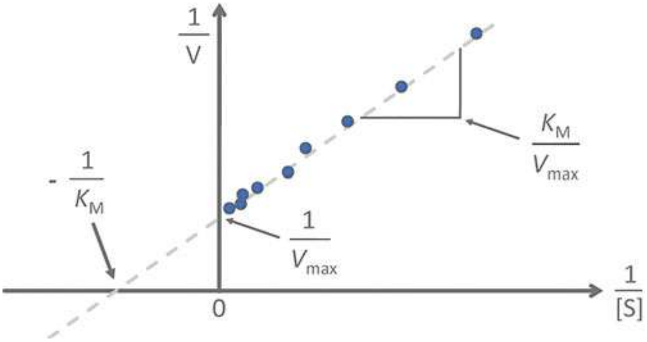


Fig. 11.13 Lineweaver-Burk diagram as a double reciprocal plot for linearizing the Michaelis-Menten equation. Measuring points can only be recorded in the positive range of 1/[*S*]

It must be emphasized that the meaning of measurements made at low substrate concentrations is distorted in the Lineweaver-Burk diagram. Therefore, it is now considered that although linear plots are useful for visualizing data, they should not be used to determine kinetic parameters, as very good computer software is now available for more accurate determination using non-linear regression methods.

11.4 Non-Michalis-Menten Kinetics

Some enzymes show a sigmoid curve instead of a saturation curve when *V* is plotted after substrate concentration. This is not consistent with Michaelis-Menten kinetics. It is a cooperative binding of the substrate to the enzyme. This means that the binding of one substrate molecule influences the binding of subsequent substrate molecules (Fig. 11.14).

This behavior usually occurs in multimeric enzymes with multiple interacting active sites and represents a regulatory mechanism. The best-known example is the binding of oxygen to hemoglobin, where the binding of oxygen to one active site alters the affinity of the other active sites for oxygen molecules. A distinction is made between positive and negative cooperativity. Positive cooperativity occurs when the binding of the first substrate molecule promotes the affinity of the other active sites for the substrate. This increases the sensitivity of enzymes to [*S*] and their activities can change greatly over a small range of substrate concentrations. Conversely, negative cooperativity occurs when binding of the first substrate decreases the affinity of the enzyme for other substrate molecules. Negative cooperativity makes enzymes insensitive to small changes in substrate concentration.

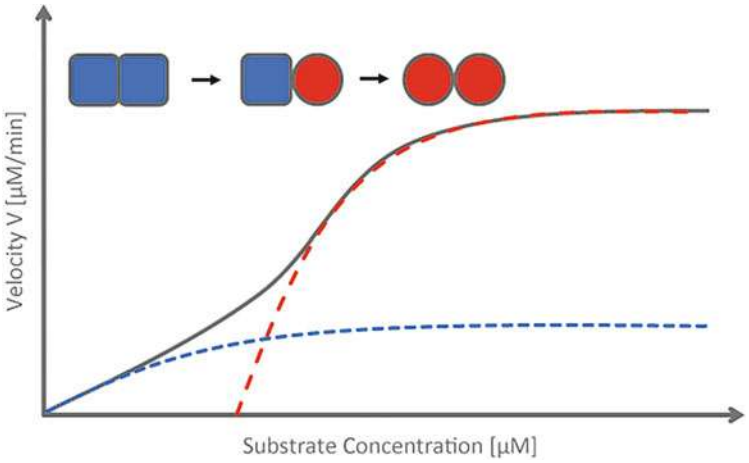


Fig. 11.14 Some enzymes show a sigmoid curve instead of a saturation curve. This indicates the cooperative binding of the substrate to the enzyme. The binding of a substrate molecule facilitates further binding. The sigmoid curve can therefore be explained by two saturation curves (red and blue dashed) of two differently active enzyme conformations

11.5 Enzyme Inhibition

It was demonstrated early on in enzymology that certain substances have a negative influence on enzyme activity. Accordingly, these are referred to as inhibitors.

By the kinetic study of enzymes in the presence of a wide variety of inhibitors, the following enzyme kinetics were observed (Fig. 11.15).

11.5.1 Reversible Inhibition

Reversible inhibition is when the inhibitor does not form an irreversible covalent bond with the enzyme and therefore competes with other molecules.

In general, there are three types of inhibitory responses, which can be classified according to the type of attack. In competitive inhibition, the target of the inhibitor is the enzyme, whereas in non-competitive inhibition the site of the attack is the enzyme and the enzyme-substrate complex. In uncompetitive inhibition, which rarely occurs, the inhibitor reacts with the enzyme-substrate complex.

Competitive Inhibition

$$\text{ATP} + \text{Glucose} \longrightarrow \text{ADP} + \text{Glucose-6-Phosphat}$$

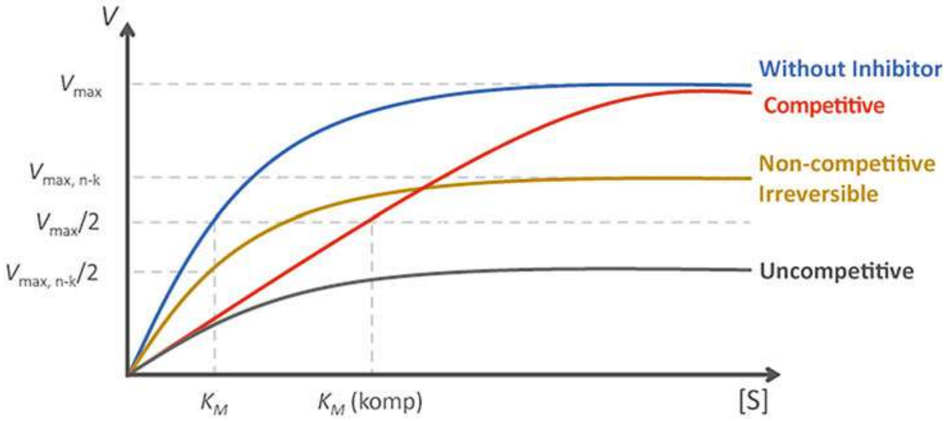


Fig. 11.15 Substrate saturation curves in the presence of different inhibitors classified into four inhibitor types (irreversible, reversible-competitive, reversible-uncompetitive, and reversible-noncompetitive)

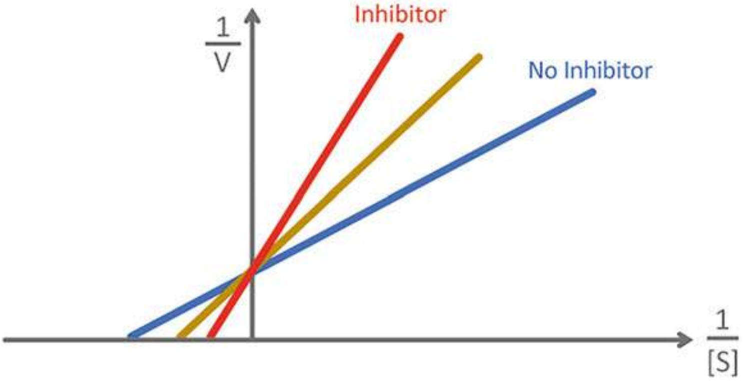


Fig. 11.16 Competitive inhibition shown in a Lineweaver-Burk diagram

The inhibitor occupies the active site and prevents the substrate from binding to the enzyme. It competes with the native substrate for the binding site. The inhibitor competes with the substrate (Fig. 11.16). Competitive inhibitors are often compounds that resemble the actual substrate or transition state.

Non-competitive Inhibition

$$\text{Glucose-6-Phosphat} + H_2O \longrightarrow \text{Glucose} + P_i$$

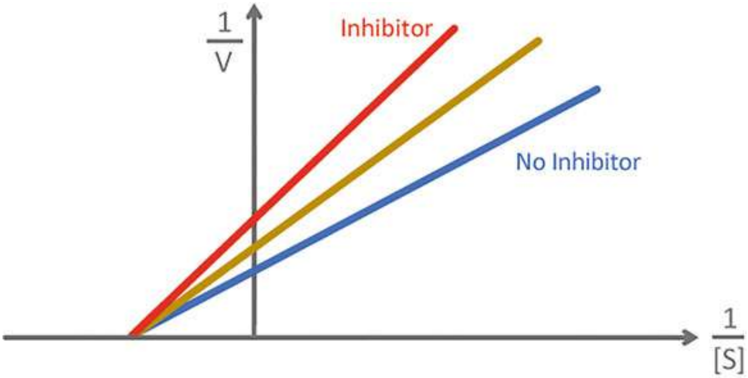
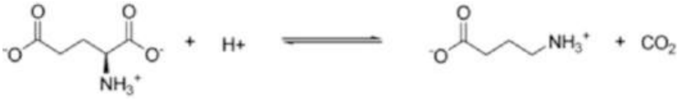


Fig. 11.17 Non-competitive inhibition shown in a Lineweaver-Burk diagram

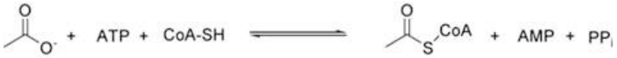
If the inhibitor binds to both the free enzyme and the enzyme-substrate complex, this inhibition mechanism is termed non-competitive (mixed) (Fig. 11.17). A non-competitive inhibitor is thought to bind allosterically to enzyme regions that do not interact with the substrate but exert an influence on the enzyme reaction.

Uncompetitive Inhibition



In uncompetitive inhibition, the inhibitor binds directly to the enzyme-substrate complex allosterically but not to the free enzyme (Fig. 11.18).

11.5.2 Irreversible Inhibition



If an inhibitor irreversibly binds to an enzyme, it is called an inactivator. Inactivators reduce the effective concentration of $[E_t]$ and therefore V_{max} without changing $K_{M.}$ The double-reciprocal plot of the rate of enzyme-catalyzed reaction against substrate concentration for irreversible inactivation is therefore similar to that for non-competitive inhibition: The straight lines intersect the $1/[S]$ axis (Fig. 11.17).

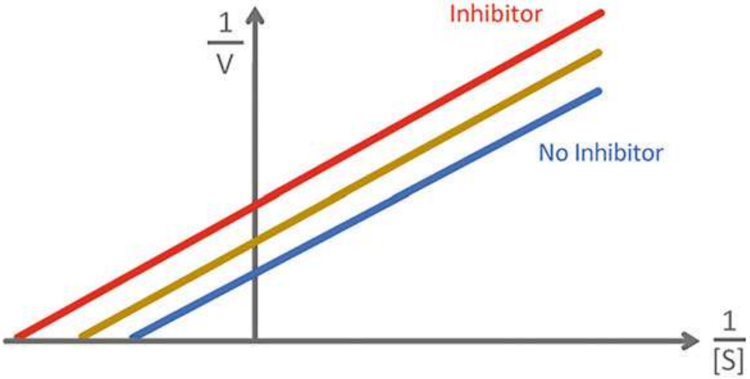


Fig. 11.18 Uncompetitive inhibition shown in a Lineweaver-Burk diagram

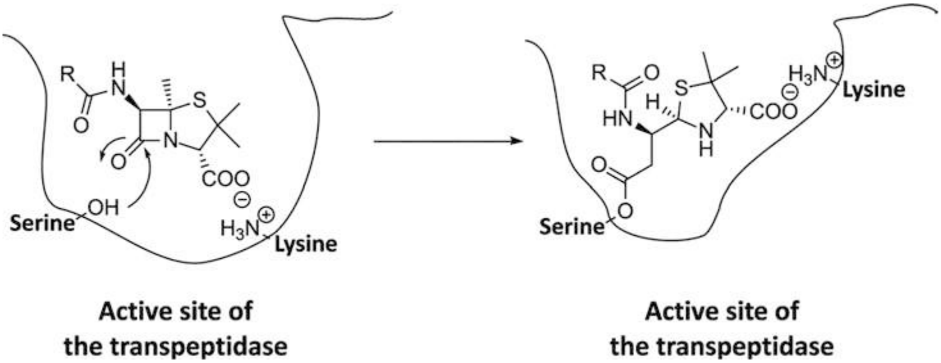

R
H
N
O
S
N
O
COO
OH
Serine
H₃N
Lysine
Active site of the transpeptidase
R
O
H
S
HN
COO
H₃N
Lysine
N
H
O
O
Serine
Active site of the transpeptidase

Fig. 11.19 Irreversible inhibition of an enzyme using the example of penicillin. Penicillin binds in the active site of the cell wall synthesis enzymetranspeptidase and binds covalently to serine. The active site is blocked and the transpeptidase is, therefore, irreversibly inactivated

The problem with many irreversible inhibitors is their strong reactivity, which is accompanied by lower selectivity for the enzyme to be inhibited. Reagents that chemically modify specific amino acid residues can act as inactivators. For example, the catalytic residues serine and histidine of serine proteases have been identified with them (Fig. 11.19).

11.5.3 Slow-Binding Inhibitors

Similar binding behavior as irreversible, covalent inhibitors is shown by so-called *slow-binding inhibitors*. In these cases, the inhibitors bind rapidly to the enzyme in a low-affinity EI complex, and this then slowly changes its conformation to a very tightly bound EI* complex (see Fig. 11.18 irreversible inhibitors). This kinetic behavior is called *slow-binding*. In summary, inhibitors that exhibit kinetics like irreversible inhibitors do not necessarily have to bind covalently to the enzyme. Examples of slow-binding inhibitors are drugs, such as methotrexate or allopurinol.

Summary

Proteins that serve as drug targets are usually receptors, transporters, or enzymes. What these three protein types have in common is that they all a) have deep pockets for the binding of a natural low molecular weight ligand or substrate (ligandability) as well as b) amplify biochemical signals (amplification) and therefore inhibition can trigger a strong clinical effect (drugability). The binding of ligands to receptors and transporters is described by the dissociation constant KD. While the binding of an enzyme inhibitor is described by its property of inhibition using the inhibition constant KI. A parallel between protein and enzyme ligands is that both can bind to the protein or enzyme in the active pocket or allosterically. Receptors, transporters, and enzymes have been the dominant drug targets in drug discovery to date. Detailed protocols exist today for the development of a drug that binds to one of these proteins, as the next chapter will show.

Further Reading

Fersht A (1999) Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding, 3rd edn. Macmillan, New York

Chemical Genomics: From Target Protein to Small Molecule Drug

12

In forward chemical genomics (Fig.12.1), agents are tested on cells to induce a biological change (phenotype). The target of the agent is then determined. In reverse chemical genomics, related protein families are screened against chemical libraries. If an active agent is found, it is tested on cells. In most cases, forward and reverse chemical genomics experiments go hand in hand in order to clearly identify the drug target.

12.1 The Concept of the Small Molecule Chemical Probe

In the past, natural substances such as penicillin, salicylic acid, or statins often formed the starting points for present-day drugs. Nature has shown that *small molecules* can be used to successfully manipulate living systems. Small molecules are characterized by their ability to cross cell membranes and bind highly specifically to a macromolecule due to their small size and partially nonpolar character. By binding to the macromolecule, its function is modulated. If the macromolecule is a DNA or an RNA molecule, the transmission of genetic information is disrupted or enhanced, i.e. the transcription of the gene is modulated. In the case of ribosomal antibiotics, translation, the synthesis of the functional carrier protein is disrupted. The modulation of proteins directly influences their function.

With the elucidation of the molecular causes of many diseases, the task facing science today, and, in particular, the discipline of medicinal chemistry is to develop *drug-like molecules* that bind highly specifically to macromolecules and thus modulate their function. But what do drug-like molecules look like? It must be ensured that the chemical probe can reach its site of action. Therefore, the molecule has to be able to interact with both an aqueous and a lipophilic environment. Only substances with intermediate lipophilicity are able to pass through both aqueous and lipophilic phases. In addition, many binding sites in macromolecules are composed of polar and nonpolar regions. Many quantitative structure-

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_12

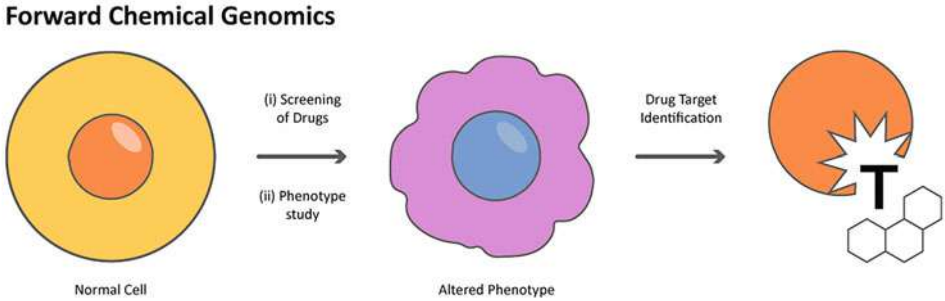


Fig. 12.1 The identification of new chemical probes and chemical probe targets is carried out by applying forward and backward genetics: (**a**) forward chemical genomics, (**b**) backward chemical genomics

activity relationships, therefore, show a clear correlation between the lipophilicity of a compound and its biological effect. The *partition* coefficient P between octan-1-ol and water has proven to be very suitable for describing lipophilicity.

$$P = \frac{[Connection]_{\mathrm{Octanol}}}{[Connection]_{\mathrm{Water}} \cdot (1-\alpha)} \tag{12.1}$$

α: Degree of dissociation of the compound in water.

For compounds that are more soluble in octan-1-ol than in water, P > 1 and log P thus becomes positive. The log P value is composed additively of the group amounts of individual parts of the molecule. Computer-aided programs already have a function for calculating log P for any compound but some of these deviate considerably from the data obtained in the experiment.

Furthermore, Christopher Lipinsky of the Pfizer company analyzed 2245 orally bioavailable chemical probes and derived general properties for these substances. The regularities that have become known as the *rule of five* (RO5) are (Fig. 12.2):

- log P < 5
- Molecular weight < 500 g/mol
- Number of hydrogen donors <5
- Number of hydrogen acceptors <10
- One of the four aforementioned rules may be violated.

However, despite the violation of the *rule of five,* many antibiotics, fungicides, vitamins, and cardiac glycosides are orally bioavailable. One explanation for this observation is that these molecules are substrates for transport systems. The *rule of five is* therefore only a guide.

Based on the observations of Christopher Lipinsky, the number of possible drug-like molecules could be calculated. As shown in Fig. 12.3, the elements carbon, hydrogen,

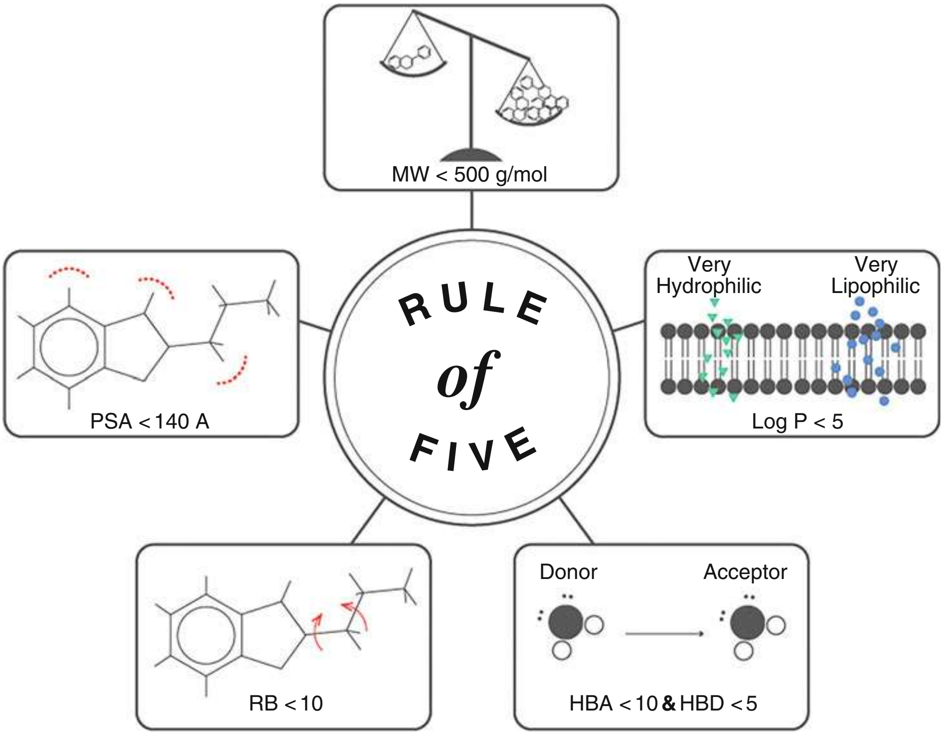


Fig. 12.2 Rule of five

nitrogen, oxygen, sulfur, phosphorus, fluorine, chlorine, and bromine, a molecular weight of less than 500 g/mol, both stability in water and less than 30 non-hydrogen atoms are present in these molecules. This gives 10^{62}–10^{63} possible molecules in chemical space (all possible synthesizable compounds, about 10^{440} compounds). It is assumed that there are about 10^{68} atoms in our Milky Way. De facto this means that all possible compounds cannot be produced due to the lack of resources (carbon atoms on earth).

12.2 Combinatorial Chemistry

Advances in molecular biology in recent years have produced a wide range of potential drug targets such as enzymes or receptors. The resulting exponential increase in demand for compounds to be tested led to the development of combinatorial chemistry in the late 1980s. This attempts to produce a large number of molecules by combining and varying different residues on a basic structure.

As shown in Fig. 12.4, combinatorial chemistry differs from the classical synthesis in that instead of two reactants A and B (classical synthesis), there are two reactant classes $\mathbf{A_n}$

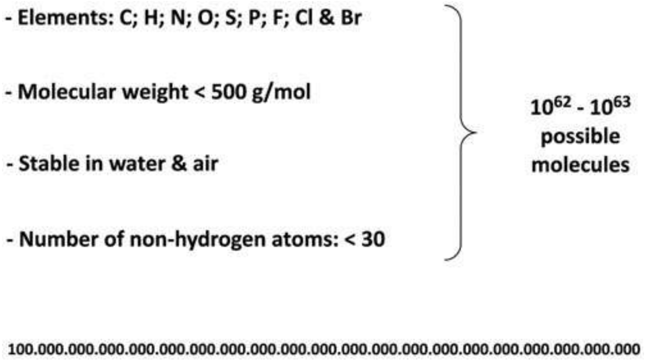

- Elements: C; H; N; O; S; P; F; Cl & Br
- Molecular weight < 500 g/mol
- Stable in water & air
- Number of non-hydrogen atoms: < 30
10^{62} - 10^{63} possible molecules
100.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000.000

Fig. 12.3 The number of all possible drug-like molecules is 10^{62} to 10^{63}

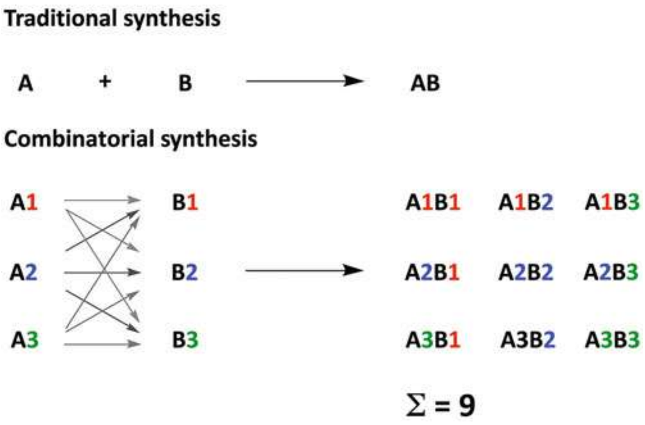


Fig. 12.4 Principle of combinatorial synthesis. (**a**) Classical synthesis, (**b**) Principle of multistep combinatorial synthesis

and $\mathbf{B_n}$ that react combinatorially. With only 3 reactants **A** and **B** each, 9 products are obtained by combination.

By combining a few building blocks, a large number of compounds, a so-called library, can thus be synthesized. If the building blocks **A**, **B** are combined with a building block **C** and 10 different compounds are used in each case, $10 \cdot 10 \cdot 10 = 10^3$ products are obtained in a very short time.

The substance libraries synthesized in this way are then tested against the drug target in biological tests called assays in *high-throughput screening* (HTS) procedures.

Since the establishment of combinatorial chemistry, many millions of compounds have been produced and tested in HTS systems. Unfortunately, it turned out that only a fraction showed biological activities. The idea of producing and testing all synthetically possible compounds (chemical space) is not feasible due to the enormous synthesis effort. Instead, the strength of combinatorial chemistry lies in the rapid and simple derivatization of a known lead structure and not, as initially assumed, in the identification of the lead structure. The problem of lead structure identification must therefore be addressed by other means.

12.2.1 Diversity-Oriented Synthesis (DOS)

With the introduction of combinatorial chemistry, it was initially believed that the number of drugs could be increased simply by a large number of synthesized and biologically evaluated compounds. Unfortunately, however, it became apparent that the first combinatorially generated libraries were not optimal, as they mostly contained large, lipophilic, flexible molecules. From this experience, efforts are now being made in the design of libraries to generate *drug-like molecules* that are as diverse as possible. This strategy in combinatorial chemistry is called *diversity-oriented synthesis* (Fig. 12.5).

12.3 High-Throughput Screening

The testing of compounds from combinatorial chemistry or diversity-oriented synthesis is automated and usually miniaturized for drug targets. The aim is to screen as many compounds as possible: *High-throughput screening* (HTS) (Fig. 12.6).

High-throughput screening is still the industry standard at the beginning of drug development or in the search for a probe in chemical biology. However, in contrast to the past, the compound libraries to be tested are more diverse in their composition or often preselected for a specific class of compounds such as kinases or GPCRs. The process is extremely capital intensive. The synthesis, storage, and testing of millions of small molecule compounds are only possible in the industry. Another problem is that the success rate of finding a hit is on average 1%, with a statistical probability of 2.5%, three times the standard deviation of the Gaussian normal distribution. In other words, statistically, the chance of finding a false positive or false negative is greater than finding a hit.

It should be noted, however, that high-throughput screening rarely finds the final drug. Rather, a hit is found that represents the starting point for further derivatives, at the end of which the lead structure is found. The lead structure is a high-affinity compound that usually requires only minor modifications.

12.3.1 Assay Design

The success of any high-throughput screening is based on the quality of the biochemical test system, the so-called assay. The quality of an assay is described, on the one hand, by its signal-to-noise ratio and, on the other hand, how close the assay parameters (pH, substrate, substrate concentration) are to the physiological conditions in the body.

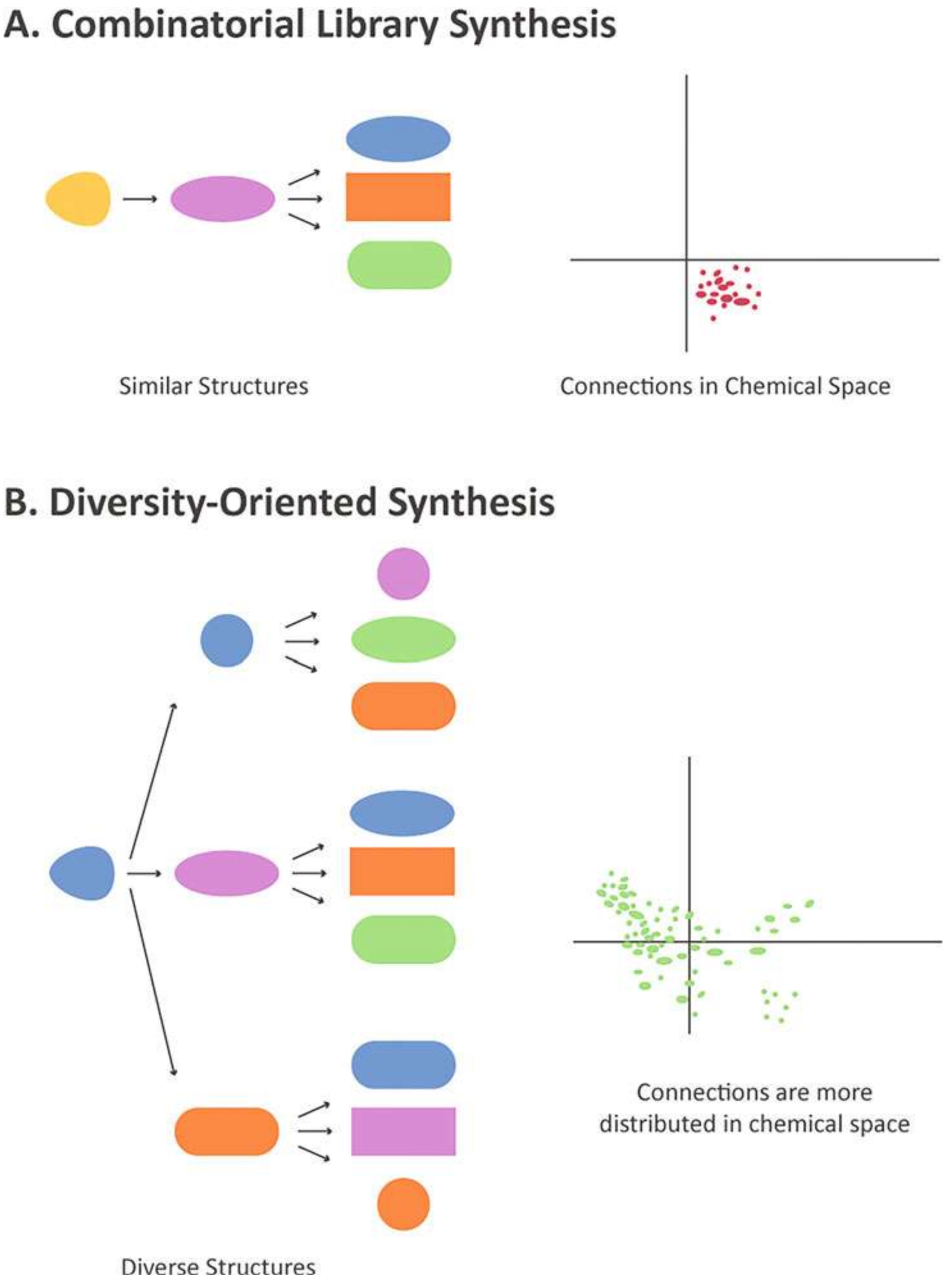

A. Combinatorial Library Synthesis
Similar Structures
Connections in Chemical Space
B. Diversity-Oriented Synthesis
Connections are more distributed in chemical space
Diverse Structures

Fig. 12.5 Combinatorial chemistry compared to diversity-oriented synthesis

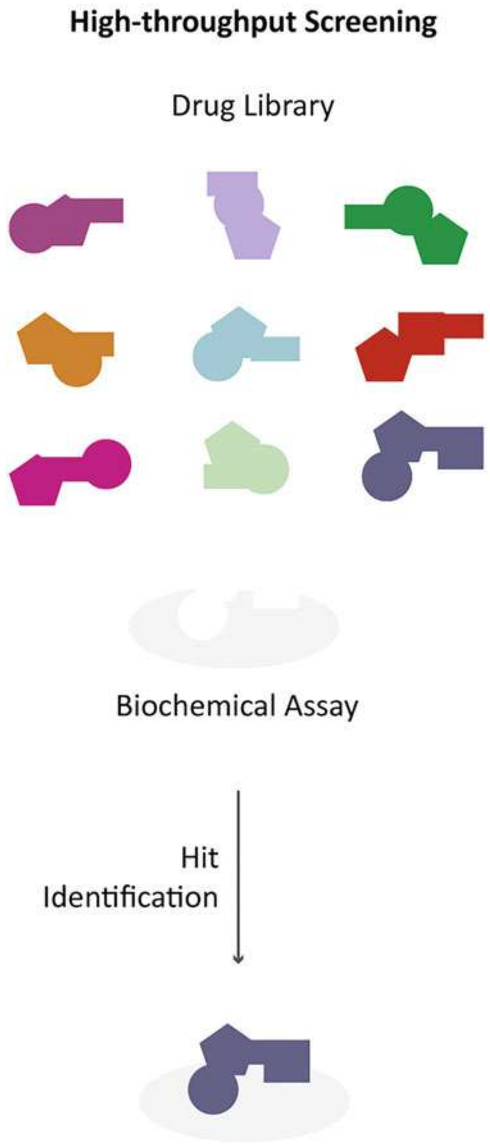


Fig. 12.6 Principle of high-throughput screening. Drug libraries of more than one million members are tested automatically in a biochemical assay for binding to the drug target or manipulation of its function

Signal to Noise Ratio

The signal-to-noise ratio can be easily determined by the Z-factor, a measure of the statistical effect size.

$$\text{Z-Factor} = 1 - \frac{3\text{SDmax} + 3\text{SDmin}}{max - min}$$

Example:

- Max − min = 100
- SDmax = SDmin = 5 → Z-factor = 0.7
- SDmax = SDmin = 20 → Z-factor = 0.2

The Z-factor should be above 0.5. An assay with a Z-factor of less than 0.5 is not reproducible.

As an alternative to the Z-factor, the assay window can be used:

$$\text{Assay Window} = \frac{\max - \min}{\sqrt{\text{SDmax}^2 + \text{SDmin}^2}}$$

Physiological Assay Parameters: One Protein: Many Drug Targets

A protein is not a statistical entity but is subject to many movements and therefore behaves dynamically in relation to its environment. The best-known example of this is the early work on kinase inhibitors in the pharmaceutical industry. Kinases are enzymes that phosphorylate serine or a tyrosine of their substrate proteins in the presence of ATP. The first kinase assays were developed using ATP concentrations in the micromolar range for cot reasons. However, the inhibitors found proved to be almost inactive in later cell assays. In the cell, the average ATP concentration is in the millimolar range, almost 1000-fold higher. The high ATP concentration causes ATP to bind to the kinase and stabilize it in the substrate-binding conformation. For inhibitors that do not bind in the substrate-binding conformation of the kinase, their drug target is not present at high ATP concentration (Fig. 12.7).

The different conformations of the protein significantly influence the binding of a ligand. Therefore, one speaks of one protein and many drug targets. The choice of assay conditions should therefore always be such that the natural conformation of the protein is present in the assay.

12.3.2 Pan Assay Interfering Substances: PAINS

Pan-assay interference substances (PAINS) are chemical compounds that frequently lead to false-positive results in high-throughput screening. PAINS bind non-specifically to numerous drug targets due to their chemical structure (Fig. 12.8). Chemical structures that allow nonspecific binding to a protein include toxoflavin, isothiazolone, hydroxyphenylhydrazone, curcumin, phenolsulfonamide, rhodanine, enone, quianone, and catechol. It is not even possible to describe here the resources that have gone into optimizing these structures in the past. For this reason, the scientific journal *Nature* 2016 described the problem under the motto “the usual suspects” (see Table 12.1).

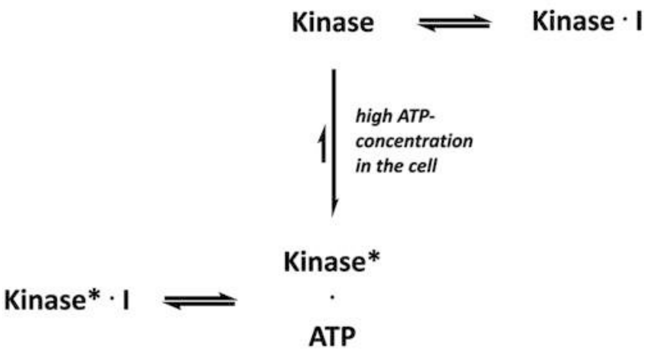
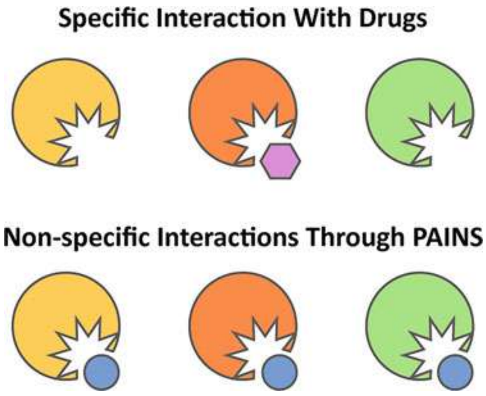


Fig. 12.7 One protein: Many drug targets

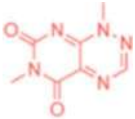
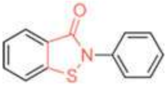
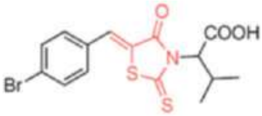
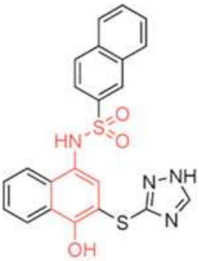
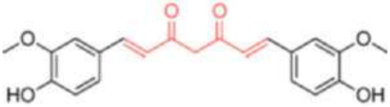
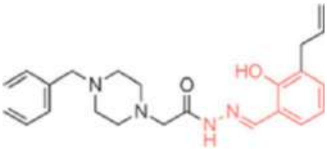
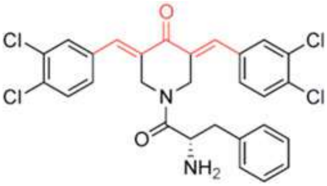
Fig. 12.8 PAINS are compounds that bind non-specifically to macromolecules due to their chemical structure


Specific Interaction With Drugs
Non-specific Interactions Through PAINS

12.4 Fragment-Based Drug Development

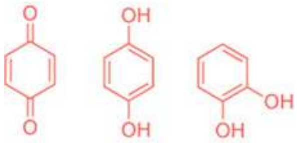
Fragment-based drug discovery was developed in the 1990s as an alternative to combinatorial chemistry and high-throughput screening. In contrast to combinatorial chemistry and high-throughput screening (Fig. 12.9a), which use large chemical libraries of more than one million substances and therefore require considerable effort, the fragment-based approach uses small libraries of less than 1000 so-called fragments, chemical compounds that, with a molecular weight of less than 300 g/mol, are significantly smaller than drug-like compounds. Conceptually, fragment-based drug discovery is based on the idea that the free energy of binding of ligand results from the individual contributions of its molecular components. The initially small contributions of the fragments can add up to form a high-affinity protein-ligand. As shown in Fig. 12.9b, one low-affinity fragment binds in each of the available binding pockets. If the two are joined, the binding energies of the low-affinity fragments add up. In addition, there is a gain in binding energy due to the reduced entropy. Two molecules have become one.

Table 12.1 The usual suspects. Structures that bind nonspecifically to proteins due to their chemical property and thus falsify results of high-throughput screening: Toxoflavin, isothiazolone, hydroxyphenylhydrazone, curcumin, phenolsulfonamide, rhodanine, enone, quianone, and catechol

Name	Structural formula	
Toxoflavin		In the presence of reducing agents such as dithiothreitol (DTT), toxoflavin forms hydrogen peroxide in the assay buffer, which oxidizes cysteines in particular and thus denatures the protein
Isothiazolone		Isothiazolones bind non-specifically covalently to the protein
Rhodan		Rhodanines bind non-specifically covalently to the protein
Phenylsulfonamide		Covalent modifier, unstable compound: Decays into molecules that provide false signals
Curcumin		Covalent modifier, membrane disruptor: Manipulate the response of membrane receptors
Hydroxyphenylhydrazone		Covalent modifier, metal complexes: Binds metal ions that inactivate proteins
Enon		Enones bind non-specifically covalently to the protein

(continued)

Table 12.1 (continued)

Name	Structural formula	
Catechol		Catechols are redox cyclers and metal ion binders, moreover, they can bind non-specifically to the protein

$$\Delta G_{AB} = \Delta G^i_A + \Delta G^i_B + \Delta G_s$$

The concept was developed in academic groups and biotech start-ups. The number of possible fragments with a molecular weight of 300 g/mol and up to 12 heavy atoms is significantly smaller at 10^7 in contrast to the 10^{63} possible drug-like substances. Common fragment libraries contain on average 1000 substances. Fragments cover the chemical space more efficiently than drug-like compounds can. Instead of testing millions of drug-like compounds against a drug target, small libraries of fragments are tested and the fragments found are subsequently elaborated into a high-affinity ligand.

As shown in Fig. 12.10, the fragment-based approach is carried out in several sub-steps. First, a thermal shift assay measures the change in thermostability of the folding of the drug target in the presence of a fragment. The thermal shift assay allows the detection of the smallest ligand bonds. Subsequently, the fragments found are validated using nuclear magnetic resonance spectroscopic techniques, such as WaterLOGSY or STD. In the next step, the binding constants of the identified fragments are determined. In contrast to high-throughput screening, questionment hits are not defined by their binding constants but by their ligand efficiency (LE). Ligand efficiency is defined as the negative quotient of the binding energy and the number of heavy atoms:

$$\text{Ligand efficiency(LE)} = -\frac{(\Delta G)}{\text{Number of heavy atoms } N}$$

As a rule, the ligand efficiency of a fragment should be 2 kJ/mol for further elaboration. Despite their low binding constants in the micro- or millimolar range, fragments with high ligand efficiency can be crystallized very well with the drug target. This is where the biggest hurdle of the fragment-based approach becomes apparent: Structural information in the form of an X-ray structure analysis is required for further elaboration. Once this is available, the fragment can be developed into a low-affinity ligand by molecular docking. The steps of molecular docking, chemical synthesis, and biological testing are carried out in an iterative cycle.

Three strategies have emerged overtime for the elaboration of a fragment into a low-affinity ligand:

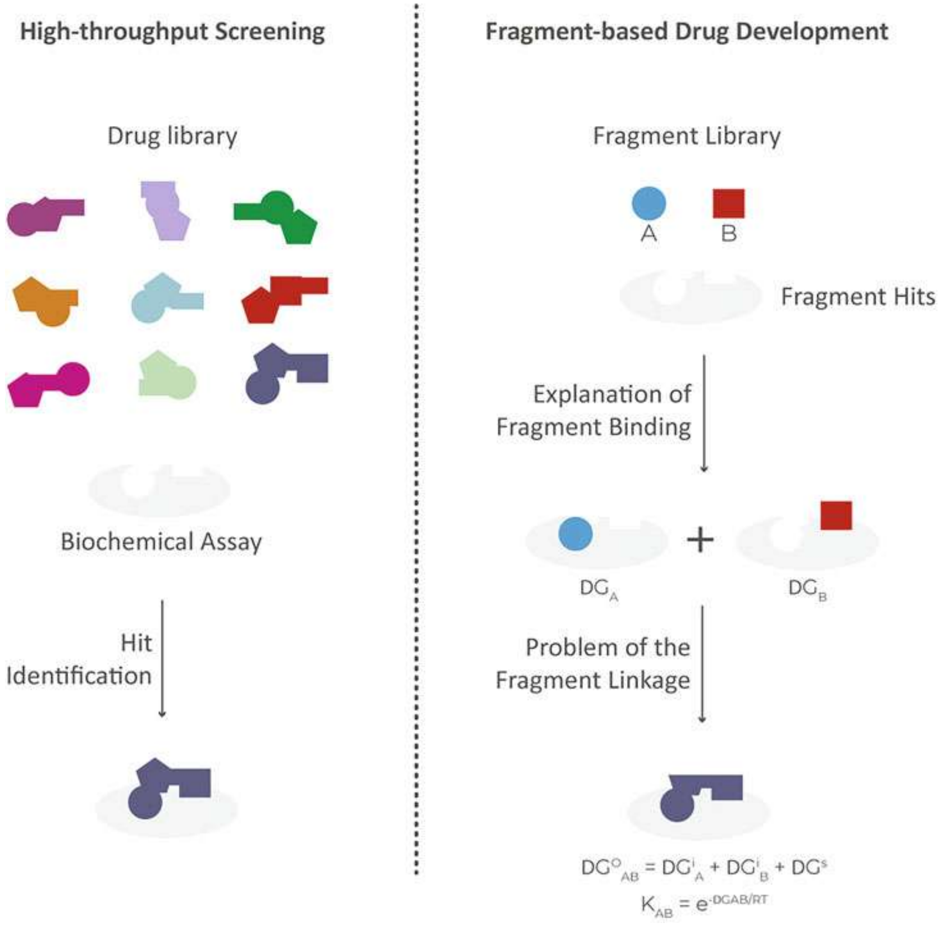


Fig. 12.9 Concepts for lead discovery. (**a**) High-throughput screening (HTS). A library of chemical compounds is tested against the drug target. (**b**) Fragment-based lead discovery. The binding of small, molecular fragments to the protein is detected. Low-affinity fragments can be linked to obtain high-affinity ligands. However, finding and linking these fragments is a difficult task. The binding energy of protein-ligand results from the additive contributions of its molecular components (A and B). The binding constant K_{AB} is an exponential function of the binding energy

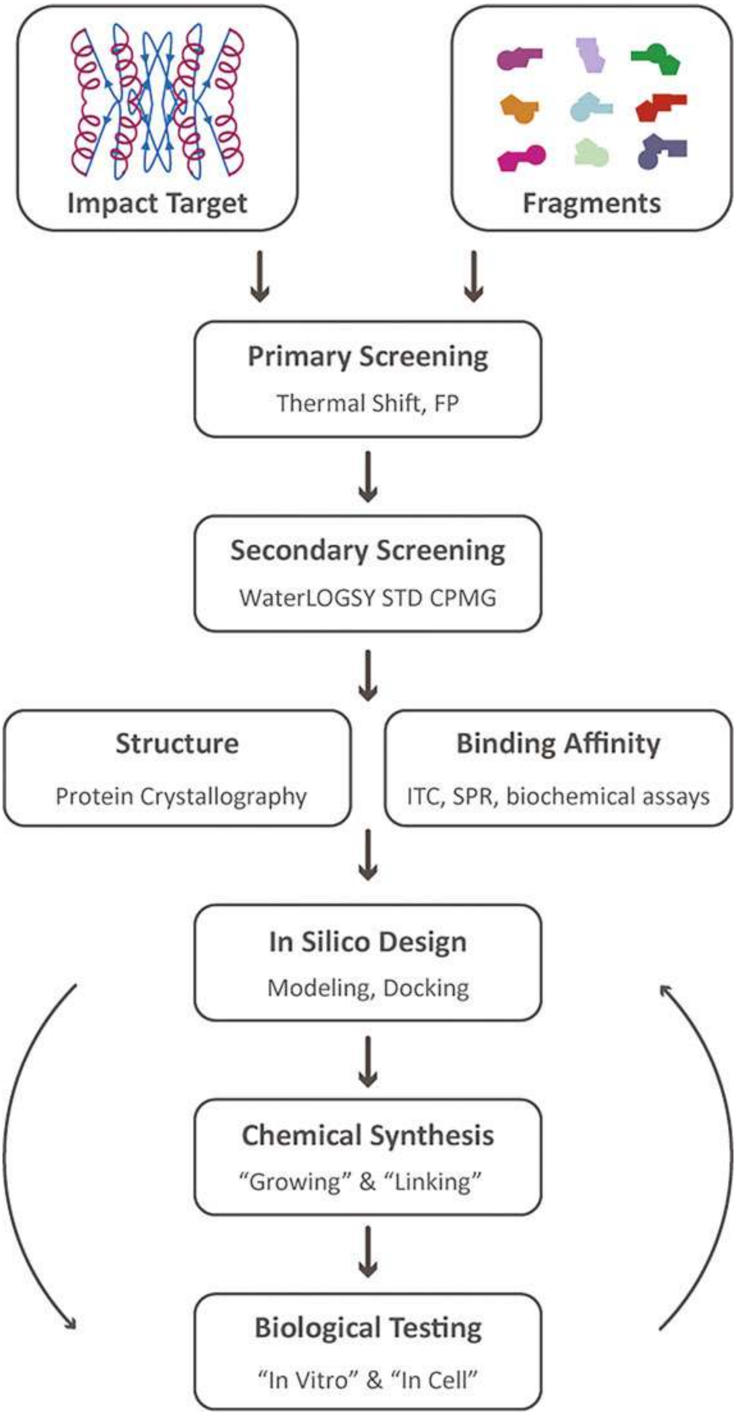


Fig. 12.10 The iterative cycle of fragment-based drug development

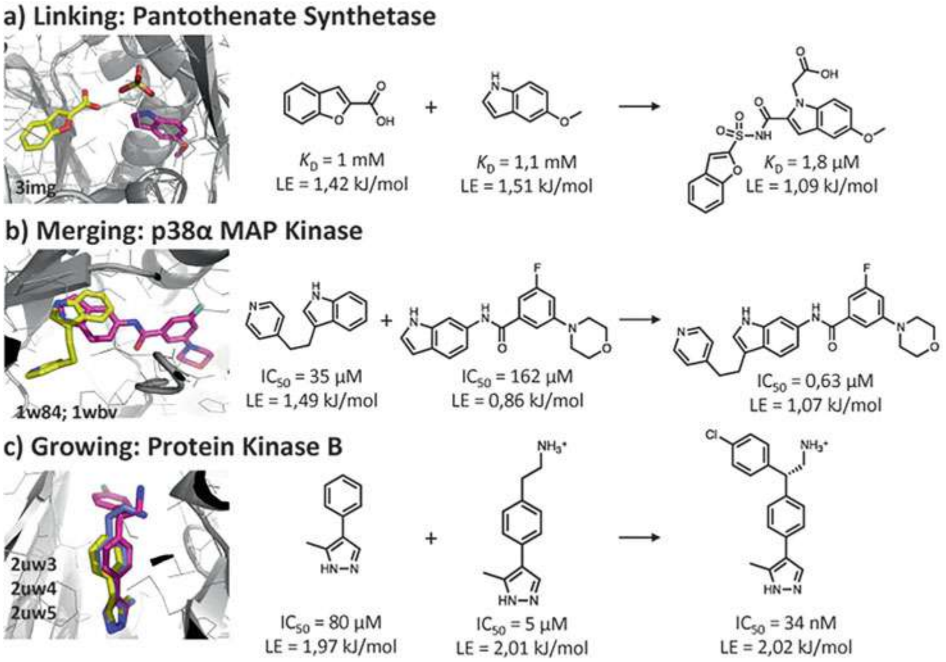


Fig. 12.11 Elaboration of a fragment: (**a**) linking, (**b**) merging, (**c**) growing

Overview

Linking: A prototype for the fragment-based approach is the linking of two fragments that bind to each other in two adjacent pockets. However, like the crystal structure is shown in Fig. 12.11a, the crystallization of two fragments in adjacent pockets is extremely rare. Fragment linking is, therefore, rather of a theoretical nature.

Merging: Merging of two fragments that form independently into the same pocket is much more common than linking. As shown in Fig. 12.11b, two fragments in two different crystal structures bind the same pocket, while having a linking indole structure. In merging, the common structural elements of the two fragments merge to form a molecule containing the different substituents of the original fragments.

Growing: Growinga fragment from one pocket to the next is perhaps the most commonly used strategy (Fig. 12.11c). In contrast to linking and merging, one only needs the *one* crystal structure of a bound fragment. As mentioned earlier, the structural information of a fragment is essential and often a crystal structure can only be obtained for one fragment.

12.5 Template-Based Drug Development

The success of the fragment-based approach stands and falls with the structural information of a bound fragment for elaboration to a low-affinity ligand (Fig. 12.12b). To be independent of the structural information for further elaboration, template-based drug discovery has evolved. Here, the fragment found is provided with a chemically reactive group and tested against other reactive fragments in the presence of the drug target as a template (Fig. 12.12c). Only fragments stabilized by the template (drug target) react with each other. To ensure that the reaction occurs under the catalysis of the template, reversible reactions such as the formation of imines from amines and aldehydes or disulfide bridges are usually used. Based on the reactive fragments and the methods used to detect the ligation products formed, template-based methods are divided into dynamic combinatorial chemistry, dynamic ligation screening, tethering, and substrate activity screening.

12.5.1 Dynamic Combinatorial Chemistry

Dynamic combinatorial chemistry (DCC) is the best-known method of template-based drug development. It is based on a reversible (dynamic) equilibrium reaction such as the formation of an imine from an amine and an aldehyde. As shown in Fig. 12.13, one incubates a known inhibitor with an amino group and a variety of ketones. In an aqueous solution, an equilibrium is formed from the amines, ketones and the imines formed. In the next step, one adds the drug target, the template, which stabilizes imines that bind better to the template. In the last step, a reducing agent is added, which converts the formed imines into stable secondary amines. The mixture obtained is measured in HPLC. Secondary amines formed only in the presence of the template and not in the control without the template are detected in LC/MS.

DCC can be applied to all types of macromolecules. It is not limited to enzymes or proteins. However, the DCC process is limited. First, one needs the drug target in stoichiometric amounts to the amines and aldehydes or ketones used. For some proteins, which are difficult to prepare, this poses a problem. In addition, larger libraries are difficult to screen because of the limitations of HPLC separation. In addition, the recognition of preferentially formed library members via reduction of the imine has not been shown to correlate with the affinity of the original ligation product. In Fig. 12.13, it can be seen that the ligation product of the compound on the lower left has an amplification rate of 84 but only an affinity of $K_I = 700$ nM. In contrast, the amplification rate of the middle compound is only 30 but it shows an affinity of $K_I = 85$ nM. Specifically, due to the changes in electronic nature from an imine to a secondary amine, one cannot infer a potent inhibitor from the ligation product.

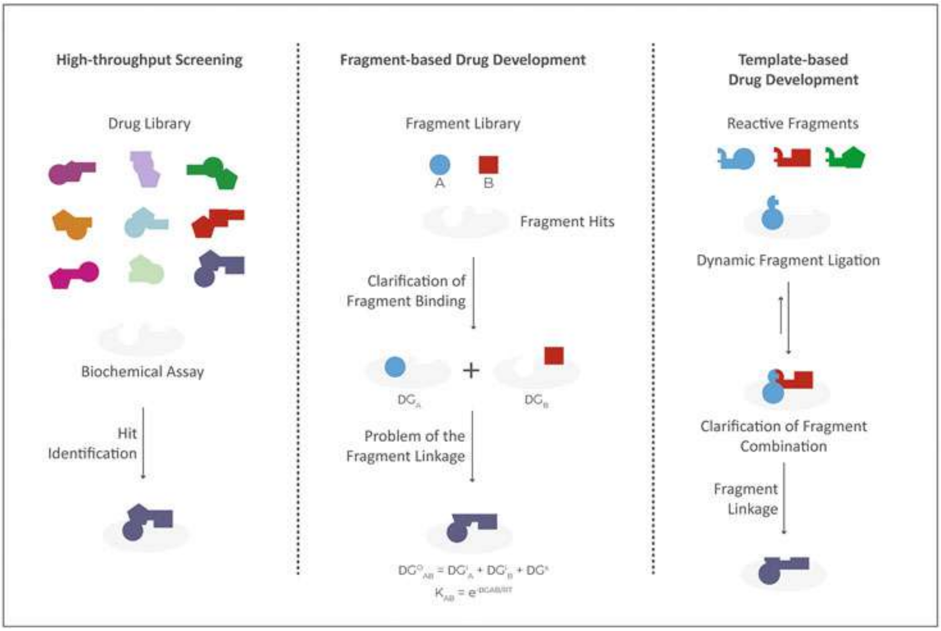

High-throughput Screening
Drug Library
Biochemical Assay
Hit Identification
Fragment-based Drug Development
Fragment Library
A
B
Fragment Hits
Clarification of Fragment Binding
DG_A
+
DG_B
Problem of the Fragment Linkage
$DG^0_{AB} = DG'_A + DG'_B + DG^*$
$K_{AB} = e^{-DGAB/RT}$
Template-based Drug Development
Reactive Fragments
Dynamic Fragment Ligation
Clarification of Fragment Combination
Fragment Linkage

Fig. 12.12 Concepts for lead discovery. (**a**) High-throughput screening (HTS). A library of chemical compounds is tested against the drug target. (**b**) Fragment-based lead discovery. The binding of small, molecular fragments to the protein is detected. Low-affinity fragments can be linked to obtain high-affinity ligands. However, finding and linking these fragments is a difficult task. The binding energy of protein-ligand results from the additive contributions of its molecular components (A and B). The binding constant K_{AB} is an exponential function of the binding energy. (**c**) Dynamic strategies in fragment-based drug discovery. Reactive fragments are incubated with the protein to form specific combinations of fragments on the protein template, which facilitates fragment detection and linkage to a new ligand

12.5.2 Tethering

Tethering represents a special form of dynamic combinatorial chemistry. In contrast to DCC, this approach uses a reversible disulfide exchange instead of imine formation as a dynamic equilibrium reaction on the surface of the protein.

As shown in Fig. 12.14, a probe in the form of an irreversible inhibitor is first bound to the drug target, in this case, the enzyme caspase-3. In the next step, the protected thiol of the probe is deprotected. The caspase-3 inhibitor adduct is then tested against a library of disulfides. A complex of caspase-3 with a bound inhibitor that has formed a disulfide with optimal residue is detected by high-resolution mass spectrometry. The fragment found is later chemically linked synthetically to the inhibitor.

The advantage of the tethering method is that a pocket can be tested specifically for low-affinity ligands. Disadvantages are, as in dynamic-combinatorial chemistry, the

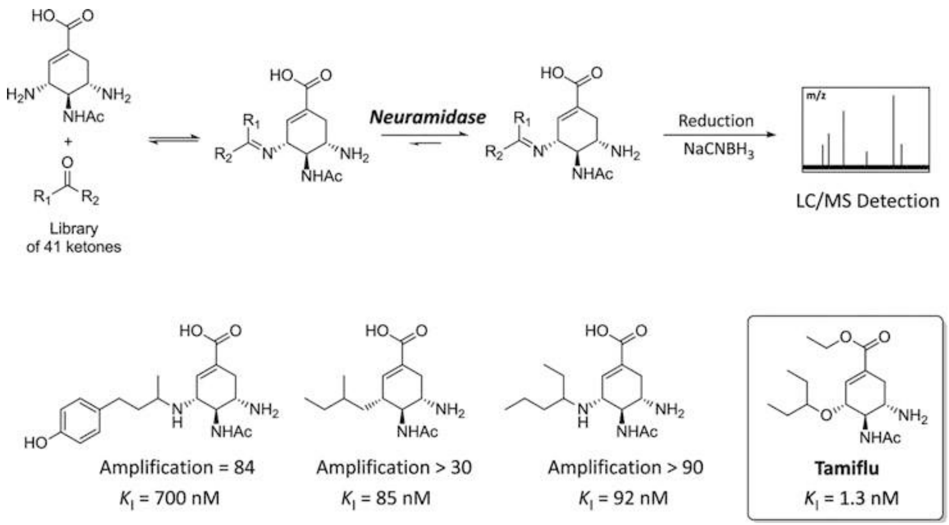


Fig. 12.13 Dynamic combinatorial chemistry. The reversible reaction of amine 1 with 41 different ketones yielded a library of 41 potential neuramidase inhibitors. Reduction followed by HPLC analysis led to the identification of several inhibitors (bottom panel). However, the compounds with the greatest enhancement did not possess the strongest affinity

stoichiometric protein consumption, the need for a special disulfide library, and the cost-intensive detection procedure using a high-resolution mass spectrometer.

12.5.3 Dynamic Ligation Screening (DLS)

Consistent further development of dynamic combinatorial chemistry is dynamic ligation screening (DLS). Similar to DCC, DLS uses imine equilibria. However, detection is not by chromatographic or mass spectrometric methods but by a biochemical assay. In particular, the use of enzyme assays to amplify the detection signal drastically reduces the amount of protein required compared to dynamic combinatorial chemistry and tethering.

The DLS approach was first applied to the major protease of severe acute respiratory syndrome (SARS) virus, a typical viral drug target similar to HIV protease. First, a fluorogenic peptide substrate is synthesized and the biochemical assay based on it is established. The SARS protease cleaves this substrate and a fluorophore is released which can be detected. Subsequently, a peptide aldehyde inhibitor is developed using the native substrate-binding sequence. The electrophilic aldehyde function forms thiolacetal with the nucleophilic cysteine residue in the reactive center of the main protease. The peptide aldehyde inhibitor is finally incubated with one nucleophilic fragment in excess at a concentration that represents its IC50 value in this assay. When a nucleophilic fragment binds to the thiol hemiacetal of peptidealdehyde inhibitor and cysteine residue of the major

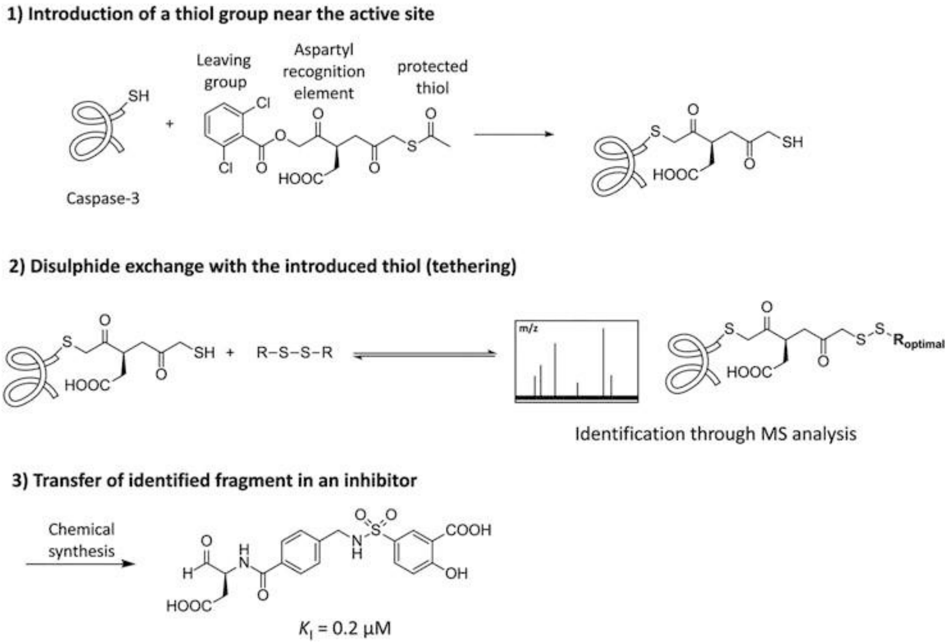

1) Introduction of a thiol group near the active site
Leaving group
Aspartyl recognition element
protected thiol
SH
Cl
Cl
O
HOOC
S
Caspase-3
SH
2) Disulphide exchange with the introduced thiol (tethering)
SH
R–S–S–R
HOOC
m/z
S-S-R_{optimal}
Identification through MS analysis
3) Transfer of identified fragment in an inhibitor
Chemical synthesis
H
HOOC
N
COOH
OH
K_I = 0.2 μM

Fig. 12.14 Example of tethering: (**a**) Introduction of a masked thiol group near the active site by crosslinking a known irreversible inhibitor targeting caspase-3. (**b**) After deprotection of the thiol, a disulfide with the highest affinity for the template stabilizes. The most stable disulfide is identified with mass spectrometric analysis of the entire complex. (**c**) Conversion of the identified fragment into a potent inhibitor of caspase-3

protease to form a full acetal, the inhibitory effect is enhanced and fluorophore formation is weaker than in the control of peptide aldehyde inhibitor alone (see Fig. 12.15).

This identifies a fragment that shows much stronger inhibition in the presence of the peptide aldehyde than the peptide aldehyde inhibitor alone. The fragment alone shows no activity towards the main SARS protease. In the final step, analogous to DCC and tethering, the peptide aldehyde inhibitor and the found fragment are synthetically linked and the new inhibitor is tested. Ultimately, one can convert the found nucleophilic fragment back into an electrophilic probe by synthetically exchanging the nucleophilic function for an electrophilic function similar to using the peptide aldehyde inhibitor used at the beginning. This time, screen a nucleophilic fragment library for hits to the pocket where the peptide structure was previously bound.

In summary, DLS provides site-specific detection of low-affinity fragments without requiring larger protein amounts like DCC and tethering, as biochemical assays are used for detection. The method can be applied especially in a high-throughput format. Unlike DCC and tethering, it does not require additional equipment such as a thiol library or high-resolution chromatographic and mass spectrometric techniques. Moreover, the size of the fragment libraries is not limited by the detection method.

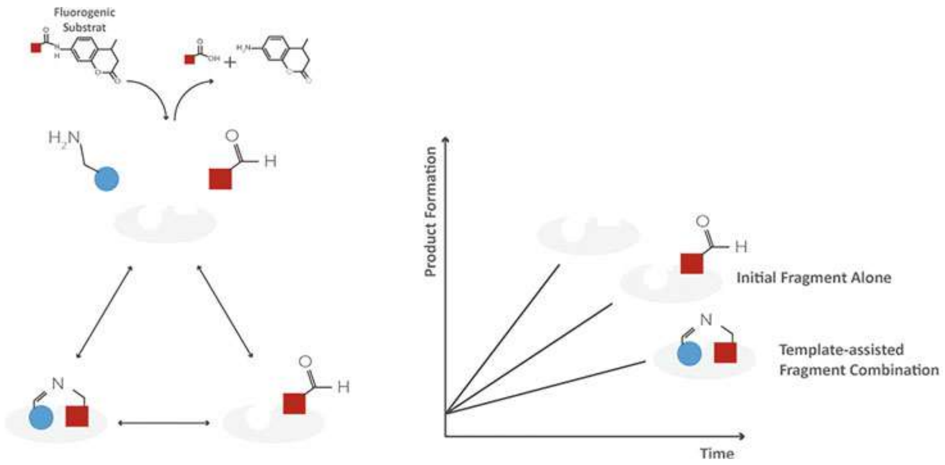


Fig. 12.15 The concept of dynamic ligation screening (DLS) using the example of a substrate-enzyme assay: The substrate competes with an aldehyde inhibitor for the binding pocket (**a**). An amine fragment binds in the adjacent pocket and simultaneously forms an imine bond to the aldehyde, resulting in increased inhibition (**b**). *Red* aldehyde inhibitor, *blue* amine fragment

12.5.4 Substrate Activity Screening (SAS)

Substrate activity screening (SAS) differs from the aforementioned methods. In contrast to the use of equilibrium reactions on the surface of the drug target, a substrate library is synthesized from fragments and tested against the drug target. The approach is limited to enzymes. As in Fig. 12.16, there are examples of proteases and phosphatases. In the case of proteases such as cathepsin S, a fluorophore is synthetically coupled to fragments via an amide bond and tested against the protease of interest by observing fluorophore formation. A high substrate turnover indicates a high affinity of the respective substrate fragments. If an active substrate fragment is found, it is synthetically optimized in the second step. The best fragment substrate is converted into an inhibitor in the last step. Here, this is done by substituting the fluorophore with an aldehyde function.

Another example of SAS is protein tyrosine phosphatase. A substrate library of O-aryl phosphates was prepared and tested against Mycobacterium tuberculosis tyrosine phosphatase B (MptpB), a possible drug target of the host-pathogen interaction of the tuberculosis bacterium. A biphenyl scaffold was identified. The phosphate group is subsequently replaced by a phosphate mimetic. After optimization, a potent and selective MptpB inhibitor is obtained.

The strength of this method is its ability to identify new substrate structures that bind to the active site. The use of protein sets in catalytic amounts contributes to the efficiency of this method. The synthesized substrate libraries can be used for all enzymes in their class. However, substrate activity screening requires the catalytic function of enzymatic drug

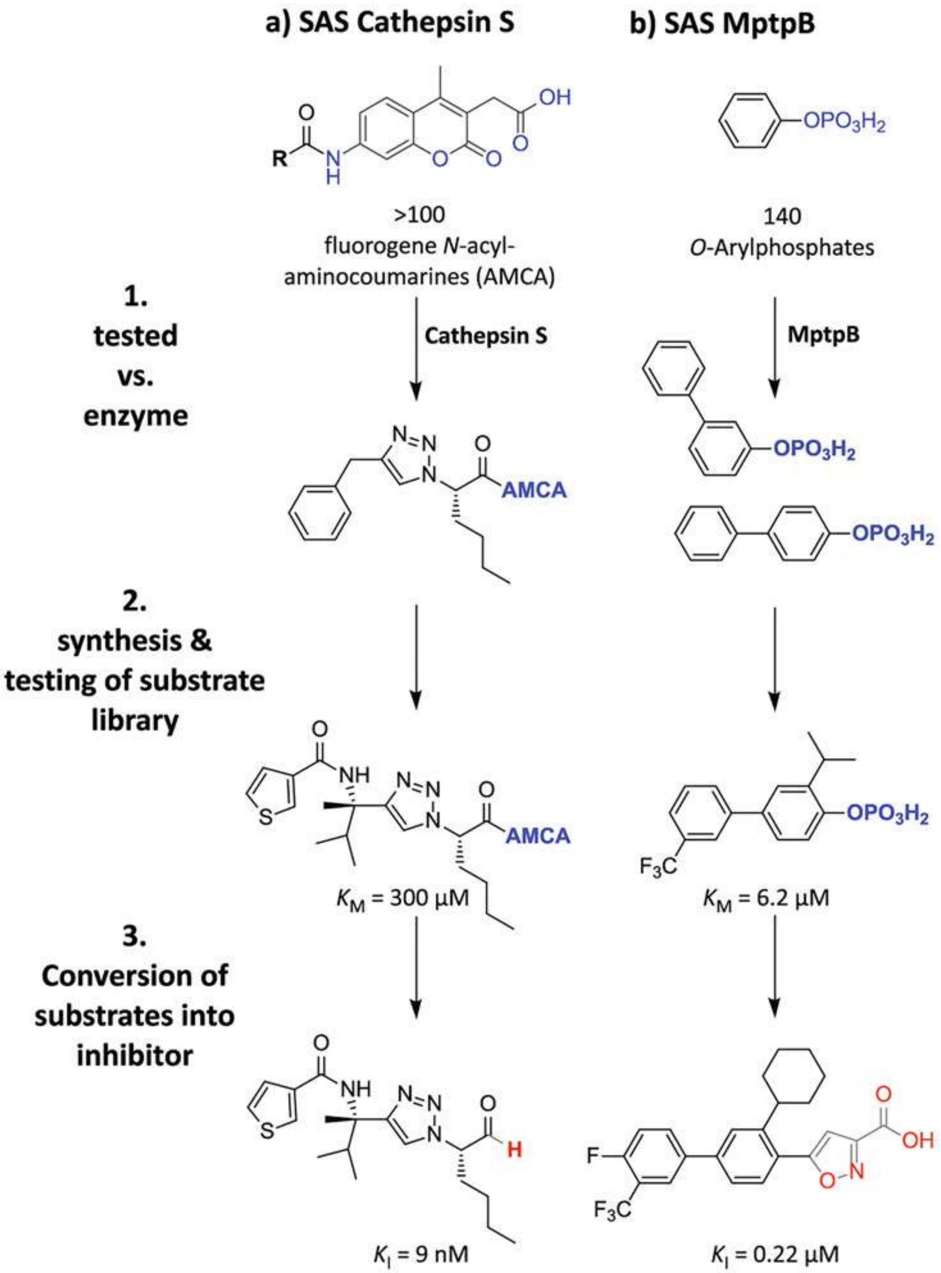


Fig. 12.16 Overview of substrate activity screening (SAS) applications. The reaction schemes for (**a**) Cathepsin S and (**b**) Mycobacterium tuberculosis protein tyrosine phosphatase B (MptpB)

targets, such as proteases or phosphatases. Another disadvantage is that the substrates found have to be converted into inhibitors similar to the other template-based methods.

12.6 Computer-Aided Drug Design

Drug design is the targeted design of a drug on the computer. The basis of drug design is the lock-and-key principle of binding a ligand to the drug target. A distinction is made between two types of drug design:

Overview

Ligand-based (indirect) drug design

Based on the shape and charge of already found ligands, new structures are predicted.

Structure-based (direct) drug design

Based on the structure of the binding pocket of the drug target determined by X-ray structure analysis or nuclear magnetic resonance spectroscopy, new structures are predicted.

Often both approaches are combined.

Drug design plays a significant role in the phase of optimizing found hits to the lead structure. In general, the following applies to computer-based methods: The more structural information of different ligands to a drug target is known, the better results can be achieved in drug design.

Excursus: Open-Source Tools Chemical Probe Design

For computer-aided drug design, a number of open-source software programs and databases have become established, which make it possible to design a drug on a home computer.

(i) **3D graphics programs for the representation of biomolecules**

The best known 3D graphics program for visualizing and rendering biomolecules is PyMOL, developed by Warren L. DeLano starting in 1998 and now distributed by Schrödinger. A limited version can be ordered from Schrödinger on the website.

https://pymol.org/2/

However, it is recommended to get a complete, open-source version via.

https://pymolwiki.org/index.php/Category:Installation

even if the installation is a bit more difficult.

(continued)

Another 3D graphics program is Chimera, developed by the University of California San Francisco (UCSF):

https://www.cgl.ucsf.edu/chimera/

(ii) **Database of 3D protein structures**

Access 165,117 protein structures (as of 11.06.2020) solved by protein crystallography or nuclear magnetic resonance spectroscopy using the RCSB Protein Data Bank (PDB).

https://www.rcsb.org/

The protein structures to be downloaded there can be opened and examined with PyMOL and Chimera.

(iii) **Determination of the 3D structure**

Structure-based drug design requires—as the name suggests—a structure. Many proteins (especially transmembrane proteins such as G protein-coupled receptors) are difficult to crystallize and the determination of the structure is therefore very limited. For proteins whose structure has not yet been solved but whose amino acid sequence is known, a homology model can be constructed by comparison with protein structures of similar amino acid sequences. The easiest way to do this is via the website of the Biozentrum of the Universität Basel SWISS-MODEL.

https://swissmodel.expasy.org/

The homology models created are now quite good and can be used for virtual screening of drug libraries or docking of specific ligands.

Even if a protein can be crystallized, the final calculation of the structure is very computationally intensive. In the meantime, there are initiatives that use crowdsourcing for these steps. One example is the computer game Fold It, in which private users can download the software from.

https://fold.it/

and solve protein structures on your home computer in a similar way to a puzzle. Another possibility is to offer the computing power of your home computer to scientists to calculate their protein structures. For this purpose, one can go to.

https://foldingathome.org/

to download software that makes the unused computing power available in the background via the Internet.

(continued)

(iv) **Virtual chemical probe libraries**

Access to or downloading of large virtual drug libraries is now possible via several internet sites:

http://zinc15.docking.org/
https://www.ebi.ac.uk/chembl/
https://pubchem.ncbi.nlm.nih.gov/

(v) **Docking programs**

Once you have the structure of your drug target via the Protein Data Bank or the creation of a homology model, and you have downloaded a virtual drug library, you can dock it against the drug target in several free simulation computer programs. The best known is DOCK from UCSF or AutoDock or AutoDock Vina from Scripps Research:

http://dock.compbio.ucsf.edu/DOCK_6/index.htm
http://autodock.scripps.edu/
http://vina.scripps.edu/

AutoDock Vina, in particular, is highly recommended for the beginner, as there is a very good tutorial in the form of a YouTube video on the aforementioned website, which guides you through the individual steps in an understandable manner.

Summary

Chemical genomics focuses on the systematic testing of gene products against low-molecular, drug-like substances for their functional manipulation and thus the possibility of developing a drug. For this purpose, established techniques from drug development are used, such as screening of large drug libraries but also the targeted development of a ligand with the help of fragment-based and template-based drug development. In this context, screening of large libraries is still the industry standard today. Nevertheless, there are already drugs that have been developed using the fragment-based approach. The template-based approach can be described as experimental and has not led to any approved drug to date.

Further Reading

Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and their exclusion in bioassays. J Med Chem 53: 2719–2740

Baell J, Walters MA (2014) Chemical can artists foil drug discovery. Nature 513:481–483

Schmidt MF, Rademann J (2009) Dynamic template-assisted strategies in fragment-based drug discovery. Trends Biotechnol 27:512–521

From Target Protein to Biologics 13

Opposite to the low-molecular active agent with its advantages and disadvantages stands the macromolecular active agent. While low molecular chemical probes can easily cross the cell membrane due to their small size and lipophilic character but always require a defined binding pocket for their binding, the macromolecular chemical probe shows the opposite advantages and disadvantages. Macromolecular drugs cannot cross the cell membrane due to their size and hydrophilic surface area. On the other hand, they do not require small defined binding pockets of the drug target, they bind to primarily extracellular drug targets over a large area of many weaker interactions, which in sum are as strong or stronger than the interactions of low molecular weight ligands. The best known macromolecular drugs are peptides, monoclonal antibodies, and aptamers.

13.1 Peptides

In contrast to proteins, peptides (up to 50 amino acids in length) do not perform functions such as enzyme catalysis but serve as hormones and growth factors in signal transduction or as antibacterial agents in immune defense. Peptides are characterized by the following properties:

Overview

i. **Peptides are the natural biological messengers in signal transduction.**

Peptides are the natural ligands for many cell surface receptors such as G-protein coupled receptors (GPCRs), ion channels, and growth factor receptors. Often peptides are agonists with a high affinity for the receptors. The best example is the

(continued)

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_13

best-known peptide drug insulin. Peptides are a good starting point for drug development due to their high affinity and selectivity.

ii. **Peptides are not membrane permeable.**

The major problem of peptidic structures in drug development is their membrane impermeability. The therapeutic application of peptides is, therefore, limited to extracellular drug targets. At the same time, the administration must be intravenous, as peptides cannot cross the intestinal mucosa. The same also applies to the blood-brain barrier. Hope is currently being raised by new forms of administration such as nanoparticles, which enable the transport of large macromolecules across the cell membrane.

iii. **Peptides are biologically unstable.**

The task of a peptide hormone is to transmit a chemical signal quickly and selectively. Therefore, peptides bind to their targets with high affinity and selectivity. At the same time, overactivation of the signaling pathway must be prevented. Peptide hormones are, therefore, degraded by proteases present. Their lifetime is very short.

Nevertheless, naturally occurring peptides can serve as starting points for drug development. The simplest example is the optimization of insulin by the introduction of further amino acids, using insulin glargine as an example, in order to increase binding to the insulin receptor (Fig. 13.1). Synthetically more complex is the development of small-molecule peptide mimetics based on natural ligands or substrates that can cross the cell membrane. Figure 13.2 shows the natural substrate of HIV protease and a drug derived from it against HIV. The introduction of lipophilic groups not only increased the affinity and selectivity of the peptide for HIV protease but also increased its ability to cross the plasma membrane. In addition, the cleavable peptide group of the substrate was replaced by an isosteroid resembling the transition state of proteolysis.

13.2 Stapled Peptides

A stapled peptide is a peptide with a synthetic “staple”. Stapling of peptides is used to make the peptides more rigid, which is associated with a decrease in entropy and thus stronger binding to the drug target (Fig. 13.3).

Small alpha-helices do not exhibit significant helicity in solution due to entropic factors. As a result, the binding affinity is reduced. The introduction of a synthetic clamp helps to fix the peptide in a specific conformation of the alpha helix, reducing conformational

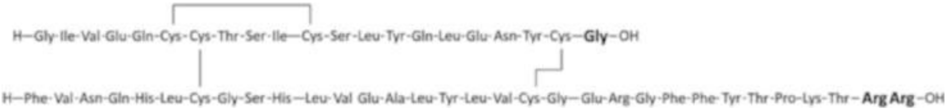

H–Gly-Ile-Val-Glu-Gln-Cys-Cys-Thr-Ser-Ile–Cys-Ser-Leu-Tyr-Gln-Leu-Glu-Asn-Tyr-Cys–Gly–OH
H–Phe-Val-Asn-Gln-His-Leu-Cys-Gly-Ser-His–Leu-Val-Glu-Ala-Leu-Tyr-Leu-Val-Cys-Gly–Glu-Arg-Gly-Phe-Phe-Tyr-Thr-Pro-Lys-Thr–Arg-Arg–OH

Fig. 13.1 Structure of the insulin derivative glargine, slightly altered in its sequence

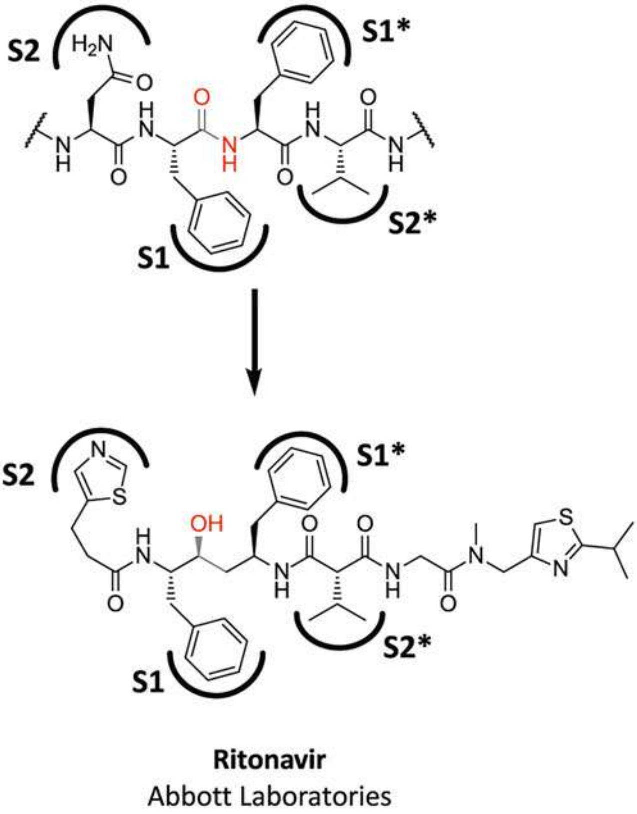


Fig. 13.2 Example of the derivation of a peptide mimetic using the HIV protease as an example: (**a**) the natural, peptidic substrate, (**b**) the inhibitor derived from it

entropy. This increases binding affinity and cell penetration and protects against proteolytic degradation.

The most common clamp is by ring closure metathesis. Using the Grubbs catalyst, ring closure occurs via two double bonds in the side chains (Fig. 13.4). The method is now well established to the extent that Fmoc-protected amino acids with the olefin side chains are commercially available and can be readily integrated into existing peptide synthesis. Due to this easy access, stapled peptides are readily used in chemical biology as probes for cell biology experiments.

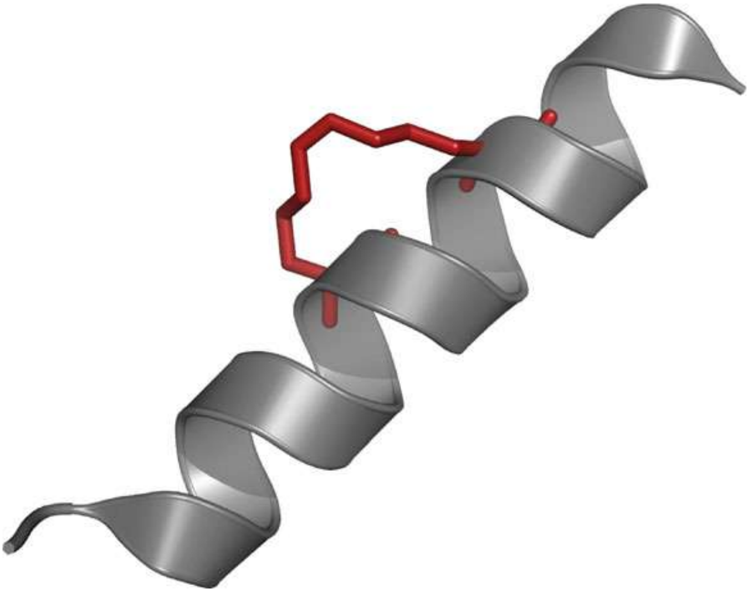


Fig. 13.3 A *stapled* peptide. Here, the alpha-helix conformation is stabilized by the staple. As a result, these peptides bind better to their targets and are protected from the natural degradation of peptides in the body

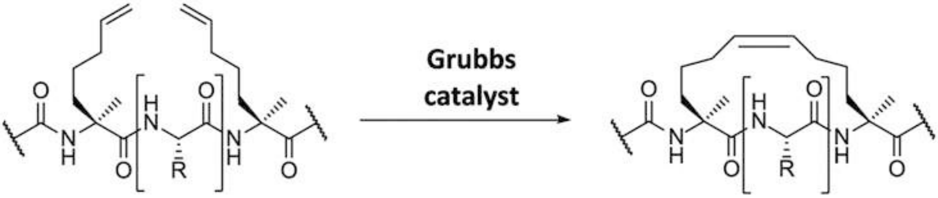


Fig. 13.4 Metastasis reaction mediated by the Grubbs catalyst to a "stapled" peptide

13.3 Monoclonal Antibodies

The largest group of macromolecular agents are the monoclonal antibodies. Antibodies, or immunoglobulins, are the key component of the adaptive immune response. They recognize foreign antigens by high-affinity binding in the picomolar range and stimulate the immune response. Structurally, the 150 kDa heterodimeric proteins consist of two heavy (50 kDa) and two light (25 kDa) polypeptide chains (Fig. 13.5). Proteolytic cleavage by the enzyme papain separates the Fab (fragment-antigen binding) portion from the Fc (fragment constant) portion of the molecule. The Fab fragments contain the variable domains

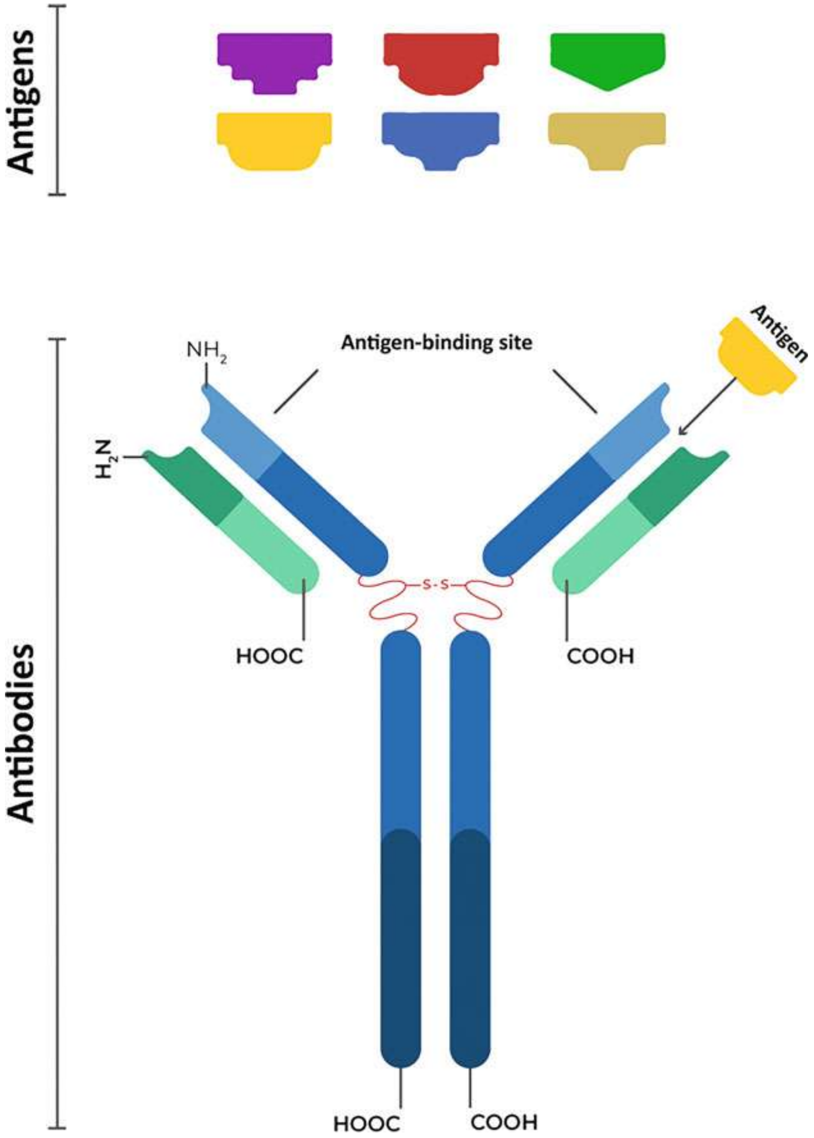


Fig. 13.5 Structure of the antibodies: The 150 kDa heavy, heterodimeric proteins consist of two heavy (50 kDa) and two light (25 kDa) polypeptide chains

consisting of three hypervariable amino acid domains responsible for antibody specificity embedded in constant regions.

The idea of using antibodies as inhibitors of extracellular proteins in medicine has been around for a long time. They help the immune system to distinguish foreign cells from body cells. Cells to which antibodies have bound are removed by the immune system.

Particularly in cancer therapy, antibodies can recognize the body's tumor cells and trigger an immune response against them. However, antibodies formed in the body are polyclonal. This means that although the antibodies bind the same antigen, they are structurally different. For the detection of a protein or other antigen in diagnostics, this initially played no role. Polyclonal antibodies are not suitable for therapeutic purposes. With a mixture of partly unknown antibody structures, a reproducible dose administration is almost impossible. In addition, the risk of side effects due to off-target effects is too high. Only the production of *monoclonal antibodies* (mAb) paved the way for antibodies as therapeutics in medicine.

The breakthrough for the production of monoclonal antibodies was the generation of hybridomas (Fig. 13.6). Here, antigen-specific plasma/plasma mast cells that produce antibodies are fused with myoloma cells. Fusion is only possible by using a medium in which only fused cells can survive. Myeloma cells lack hypoxanthine-guanine phosphoribyltransferase (HGPRT), an enzyme needed for the synthesis of nucleic acids. Usually, this is not a problem. Only if the purine synthesis pathway is blocked, for example, by the administration of the inhibitor aminopterin—a folic acid analog that inhibits dihydrofolate reductase (DHR)—do the myeloma cells need another source of nucleic acid to survive. The medium used for fusion is called HAT medium because it contains hypoxanthine, aminopterin, and thymidine. Only myeloma cells that fuse with one of the plasmablast cells can survive in this medium. These fusion cells are called hybridomas. Non-fused myeloma cells cannot grow because they lack HPGRT and therefore cannot replicate their DNA. Non-fused splenocytes cannot grow indefinitely due to their limited lifespan. Only fused hybrid cells, called hybridomas, can grow indefinitely in the medium because the spleen cell partner provides HGPRT and the myeloma partner has properties that make it immortal (similar to a cancer cell). After the non-fused spleen and myeloma cells have died, the cell suspension is diluted and aliquoted until a single antibody-producing hybridoma cell is obtained. This single cell is allowed to proliferate again until a cell culture is obtained that continuously produces only one structurally identical—monoclonal—antibody.

Therapy with monoclonal antibodies has proven particularly effective in cancer and autoimmune diseases. Monoclonal antibodies help the immune system distinguish foreign cells from body cells. Many tumor cells that proliferate at high rates, or body cells that die and subsequently cause physiological problems are not attacked by the immune system because tumor cells are the body's cells. However, tumor cells are highly abnormal and many have unusual antigens. Some such tumor antigens are inappropriate for the cell type or its environment. Monoclonal antibodies can bind very specifically to tumor cells or to abnormal cells in the body that are recognized as body cells but are detrimental to health. In reference to the concept of magic bullets that always hit the target—adapted by Paul Ehrlich from the opera "Der Freischütz" ("The Freeshooter")—monoclonal antibodies are referred to as the magic bullets of today's medicine.

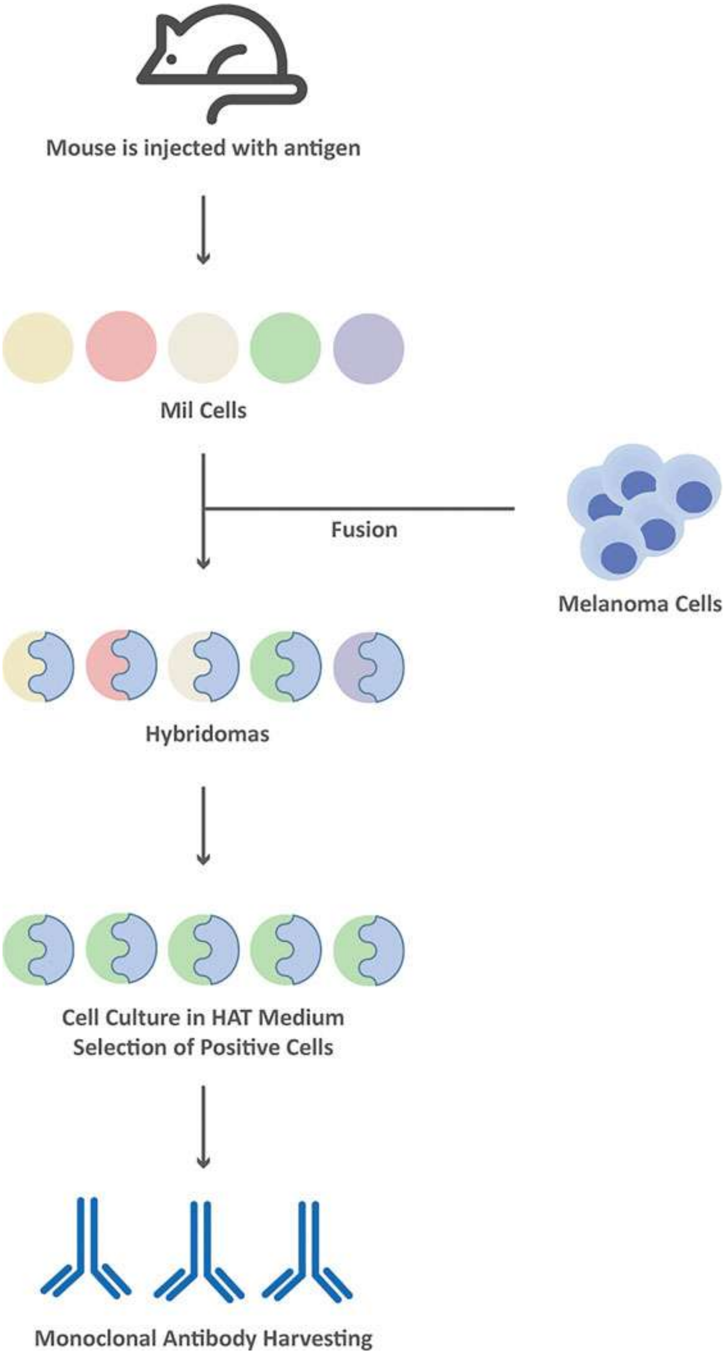

Mouse is injected with antigen
Mil Cells
Fusion
Melanoma Cells
Hybridomas
Cell Culture in HAT Medium
Selection of Positive Cells
Monoclonal Antibody Harvesting

Fig. 13.6 Schematic representation of the development of a monoclonal antibody

Excursus: Biosimilars and Generics—Wine and Lemonade

Drugs based on small-molecule compounds for which patent protection has expired and which are therefore also manufactured by other pharmaceutical companies and sold at significantly lower prices are referred to as generics. In contrast, drugs based on monoclonal antibodies, for which patent protection has also expired and which are now manufactured and sold by other pharmaceutical companies, are referred to as biosimilars.

Unlike chemical agents, monoclonal antibodies are very difficult to mimic due to their molecular complexity. The development of generic drugs costs between US$2 and 3 million, while the cost of developing a biosimilar is up to US$200 million. Biosimilar manufacturers know the structure of the monoclonal antibody from patents and publications. However, they do not have access to the cell line used and do not know the manufacturing process in detail. Ultimately, it must be proven that the biosimilar produces the same clinical effect in clinical trials as the original drug. This makes the path to the biosimilar long and rocky compared to the generic drug.

The differences between a generic and a biosimilar are often compared to the production of lemonade and wine. Lemonade is always made from the same flavoring powder. It does not matter from which production site it comes. In contrast, wines from two different wineries can never be identical. Differences exist, for example, in the grapes due to the weather, in the yeast strain used, and the pressing process.

Despite this enormous effort to develop a biosimilar, the costs are worth it. Monoclonal antibody therapies cost five to six figures per treatment. High-profit margins, therefore, exist even for biosimilars.

13.4 Aptamers

The word aptamer is derived from the Latin “aptus”—fit and “meros”—area. Aptamers are usually DNA or RNA oligonucleotides with a length of 25 to 70 bases that specifically bind small molecules, proteins, or whole cells non-covalently due to electrostatic interactions, hydrophobic interactions, and especially their complementary 3D structure. Aptamers have some advantages over antibodies in that they can be readily prepared by chemical synthesis, are generally stable at room temperature, and produce little or no immunogenicity in therapeutic applications.

Aptamers are developed using **s**ystematic **e**volution of **l**igands by **ex**ponential Enrichment—SELEX. SELEX is a combinatorial method for the targeted evolution of binding oligonucleotide strands to a selected target.

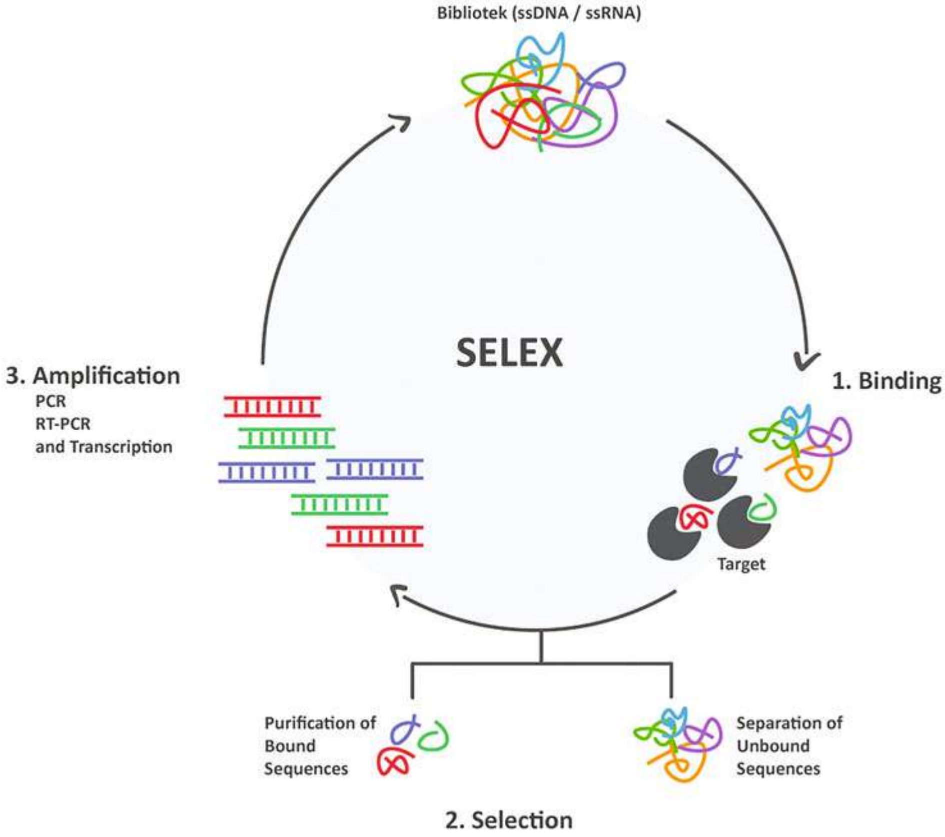


Fig. 13.7 The development of high-affinity oligonucleotides using the SELEX process

In the first step, a library of up to 10^6 single-stranded DNA or RNA oligonucleotides is incubated with the target molecule (Fig. 13.7). In the next step, the oligonucleotides bound to the target molecule are separated from the unbound oligonucleotides. The oligonucleotides previously bound to the target molecule are amplified in the next, third step by the polymerase chain reaction (PCR). Usually, a mutation step is added here by error-containing PCR to obtain possibly better binders for the next binding process. The three steps are repeated as required until a high-affinity aptamer is obtained.

Spiegelmere

The disadvantage of aptamers is their lack of resistance to nucleases and thus their low bioavailability in the body. In addition to the incorporation of nuclease-stable building blocks (see Chap. 9), the use of L-ribonucleic acid instead of the naturally occurring D-ribonucleic acid has become established. L- and D-ribonucleic acids behave to each other like image and mirror image. For this reason, aptamers consisting of L-ribonucleic acids are called mirror mers.

However, the SELEX method is not applicable to mirror mers due to the PCR used. Therefore, the target protein is often synthesized from D-amino acids instead of natural L-amino acids and then tested against an oligonucleotide library by SELEX. The oligonucleotides found are then re-synthesized from L-ribonucleic acids and tested as mirror mers against the natural protein from L-amino acids.

Summary

In addition to the small molecules, the biologics form the second large group of chemical probes and drugs. Biologics are macromolecules, such as modified peptides, monoclonal antibodies, or aptamers, and are neither orally available and therefore administered parenterally nor can they usually penetrate the cell membrane and therefore bind to targets outside the cell, such as receptors. Monoclonal antibodies have been enormously successful in recent years. Developing a monoclonal antibody against a protein offers more chance of success than finding a small molecule binder. Ultimately, the big problem remains that biologics can only address drug targets outside the cell.

Further Reading

Henninot A, Collins JC, Nuss JM (2018) Thew current state of peptide drug discovery: back to the future? J Med Chem 61:1382–1414

Köhler G, Milstein C (1975) Continous cultures of fused cells secreting antibody of predefined specificity. Nature 256:495–497

Lau YH et al (2014) Peptide stapling techniques based on different macrocyclization chemistries. Chem Soc Rev 44:91–102

Mills DR, Peterson RL, Spiegelman S (1967) An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. Proc Natl Acad Sci U S A 58:217–224

14 Chemical Proteomics: From Chemical Probe to Target Protein

Affinity chromatography plays a central role in chemical proteomics. As shown in Fig. 14.1, the chemical probe to be investigated is immobilized on a tag via a linker. Cell lysate is added over the immobilized phase. In the next step, the immobilized phase is washed. Only the protein-bound via the chemical probe remains. By cleavage of the linker, the protein falls off the immobilized phase. The so-called eluate is separated in gel chromatography and subsequently characterized by mass spectrometry.

14.1 Affinity Chromatography

There are now two methods of affinity chromatography in chemical proteomics: *Compound-centric chemical proteomics* (CCCP) and *activity-based probes profiling* (ABPP) (Fig. 14.2).

Ligand-centered chemical proteomics corresponds to the already explained classical drug affinity chromatography. The ligand to be investigated is synthetically provided with a linker and bound to a solid phase, usually small resin beads. Cell lysate is placed over the resin beads with the bound ligand. Proteins that bind the ligand bind over it to the resin beads. In the next step, the unbound proteins are washed away. In the final step, the binding of the ligand to the resin bead is released or eluted with a highly concentrated solution of free ligands. The eluate is analyzed by mass spectrometry. The amino acid sequence of the protein found can then be determined via the mass or the decay of the protein in the mass spectrometer.

In contrast, the activity-based probe profile differs in that the probe is an irreversible inhibitor that mechanistically binds to all members of the same protein family. Here, the goal is not to find the unknown target but to investigate the profile of the occurring family members of the drug target. Hence, the name refers to “profiling”. The probe, therefore,

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_14

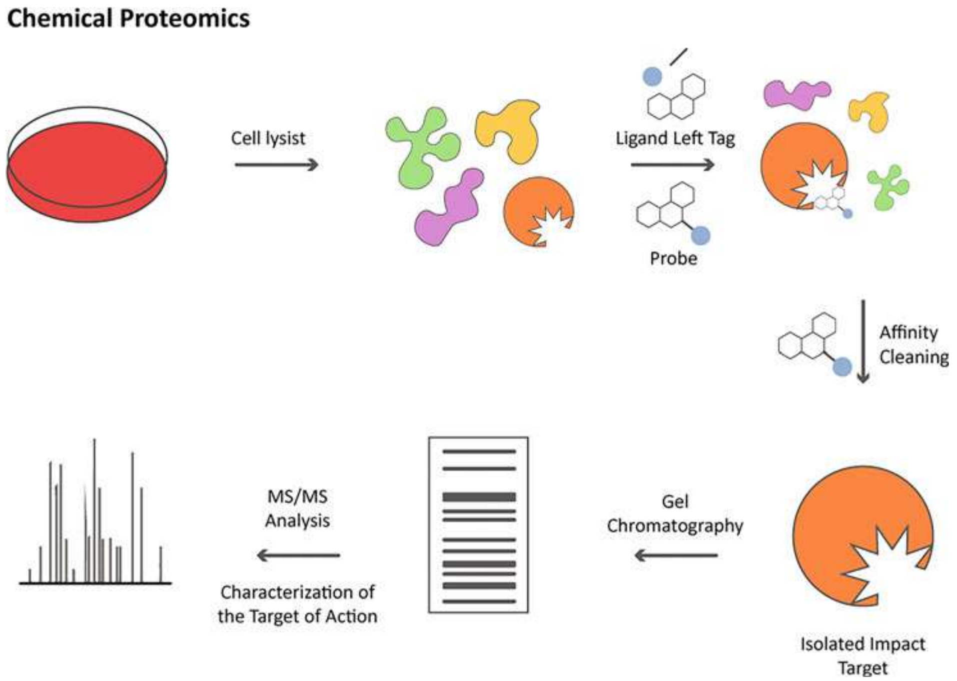
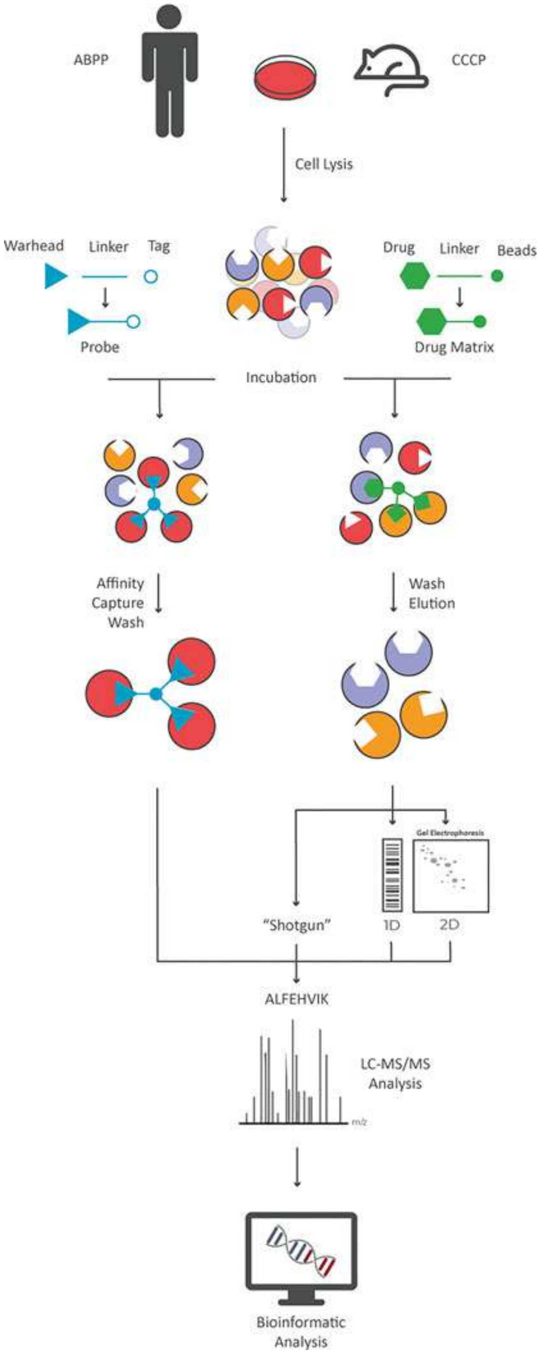


Fig. 14.1 The clear identification and characterization of the ligand-protein complex is the goal of chemical proteomics. For this purpose, the ligand is chemically labeled (tagged), which allows the target protein of the ligand to be isolated via affinity chromatography. Subsequently, the target protein is examined by gel chromatography and mass spectrometry

consists of a reactive group and a label ("tag"). The reactive group usually contains an electrophile that can covalently bind to a nucleophilic residue in the active site of an active enzyme. An enzyme that is inhibited or post-translationally modified will not react with the activity-based probe. The label can be either a reporter such as a fluorophore or an affinity label such as biotin, or an alkyne or azide to use 1,3-dipolar Huisgen cycloaddition (also known as click chemistry) (Fig. 14.3). The advantage of ABPP is the ability to directly detect enzyme activity rather than being limited to protein or mRNA abundance. With classes of enzymes, such as serine and metalloproteases that often interact with endogenous inhibitors or exist as inactive zymogens, this technique offers a valuable advantage over conventional techniques that rely on both the presence of the enzyme and its activity.

Fig. 14.2 Comparison of (**a**) activity-based probe profile (ABPP) with (**b**) ligand-centered chemical proteomics (CCCP)



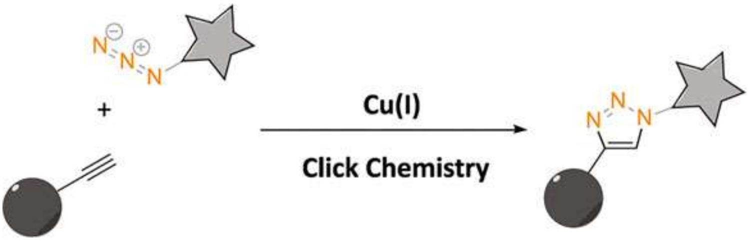


Fig. 14.3 The term "click chemistry" was coined by Nobel laureate K. Barry Sharpless to describe the copper-catalyzed [3 + 2] cycloaddition between alkynes and azides. The two components—azido and alkyne groups—are almost never found in natural biomolecules. Conjugation reactions, therefore, achieve a remarkable degree of selectivity and specificity

14.2 Polypharmacology

In classical pharmacology, it is assumed that a drug binds to a drug target. However, many affinity chromatography experiments have shown that a chemical probe rarely binds to only one drug target. The fact that a chemical probe has several chemical probe targets at the same time is referred to as polypharmacology (Fig. 14.4).

Polypharmacology is currently the subject of critical debate. On the one hand, the binding of a chemical probe to many proteins is regarded as a major risk of triggering off-target effects and thus side effects. On the other hand, it is hoped that more effective drugs can be developed through targeted modulation of multiple drug targets. It is generally believed that complex diseases, such as cancer and central nervous system diseases require complex therapeutic approaches. In this regard, a drug that "hits" multiple sensitive nodes belonging to a network of interacting drug targets offers the potential for greater efficacy and may limit the drawbacks that generally result from using a single drug or a combination of multiple drugs.

It should be noted here that chemical biology has been accused in this context of considering chemical probes only as highly selective small molecules that allow the modulation and study of a specific target, ignoring the possibility that a probe can bind to multiple proteins.

Summary

The goal of chemical proteomics is to identify and characterize the drug-drug target complex. Two affinity chromatography strategies have been established for this purpose: Activity-based probe profiling (ABPP) and ligand-centric chemical proteomics (CCCP). CCP corresponds to classical drug affinity chromatography, whereas ABPP is characterized by the fact that the probe is an irreversible inhibitor that binds

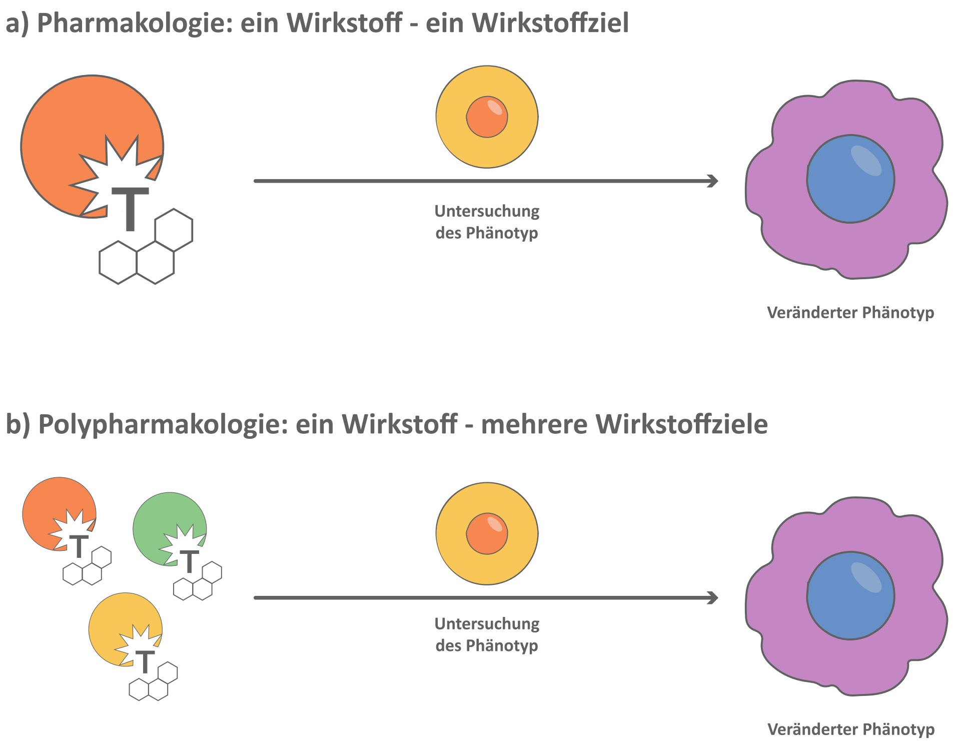

a) Pharmakologie: ein Wirkstoff - ein Wirkstoffziel
T
Untersuchung
des Phänotyp
Veränderter Phänotyp
b) Polypharmakologie: ein Wirkstoff - mehrere Wirkstoffziele
T
T
T
Untersuchung
des Phänotyp
Veränderter Phänotyp

Fig. 14.4 The difference between pharmacology and polypharmacology: (**a**) in classical pharmacology, a drug binds to one drug target. (**b**) While polypharmacology takes into account the fact that many drugs only trigger a clinical effect by binding to multiple drug targets

mechanistically to all members of the same protein family. In addition to characterizing the drug-drug target complex, chemical proteomics offers to explore the polypharmacology of drugs. What other proteins does the drug bind to? And what influence does binding to other proteins have on the clinical effect?

Chemical Genetics: Validation of the Drug Target by Elucidation of the Signaling Pathway 15

As explained in Sect. 4.3, chemical genetics is the oldest subdiscipline of chemical biology and has long been synonymous with chemical biology.

Chemical genetics is comparable to mutagenesis experiments in genetics, the only difference being that the change in the function of the gene is brought about reversibly with an agent instead of irreversibly by mutation of the gene (Fig. 15.1). Therefore, chemical genetics can be considered analogous to classical genetic screening. Here, random mutations are introduced into the organism, the phenotype of these mutants is observed, and finally, the specific gene mutation (genotype) that produced this phenotype is identified. In chemical genetics, the phenotype is not altered by the introduction of mutations but by exposure to chemical probes. Similar to chemical genomics, experiments are done in a forward and backward manner. Phenotypic screening (forward) of chemical libraries is used to find an agent that can alter the phenotype. Backward chemical genetics plays a special role in validating drug targets in experimental disease models (*target validation*): Does manipulating this protein have the desired clinical effect?

It should, therefore, be explicitly stated that the experimental setup in chemical genetics does not differ from that in chemical genomics. In chemical genetics, the focus is on validating the drug target and elucidating the signaling pathway. In order to achieve the desired clinical effect in humans, it must first be shown that this effect is detectable in cell experiments.

The use of agents to manipulate genes analogous to mutagenesis experiments in terms of chemical genetics has proven to be an integral part of elucidating signaling pathways. The best examples are thalidomide and malformation in embryos as well as FK506 and the Ca^{2+}-calcineurin pathway.

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_15

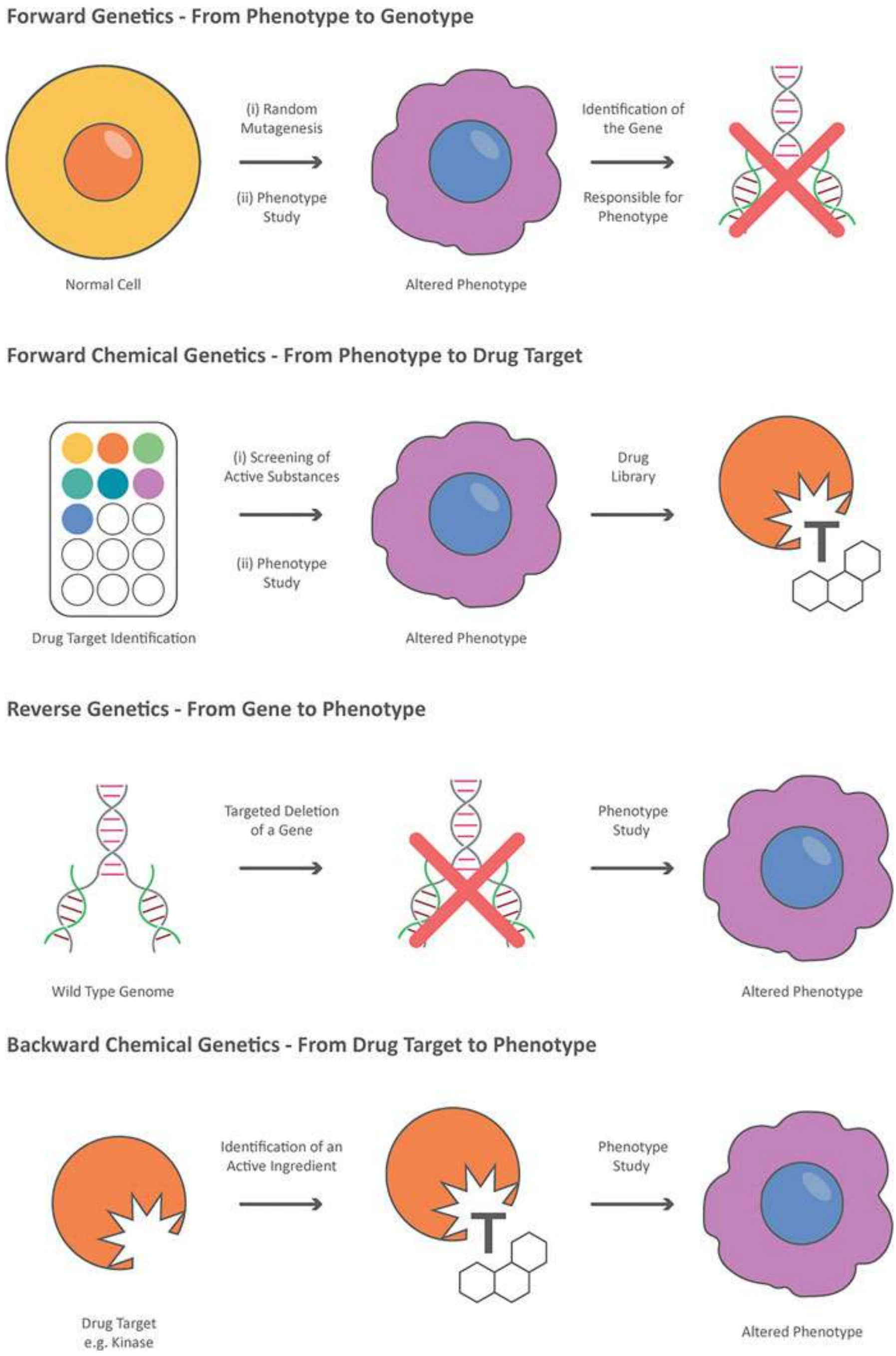

Forward Genetics - From Phenotype to Genotype
Normal Cell
(i) Random Mutagenesis
(ii) Phenotype Study
Altered Phenotype
Identification of the Gene
Responsible for Phenotype
Forward Chemical Genetics - From Phenotype to Drug Target
Drug Target Identification
(i) Screening of Active Substances
(ii) Phenotype Study
Altered Phenotype
Drug Library
T
Reverse Genetics - From Gene to Phenotype
Wild Type Genome
Targeted Deletion of a Gene
Phenotype Study
Altered Phenotype
Backward Chemical Genetics - From Drug Target to Phenotype
Drug Target e.g. Kinase
Identification of an Active Ingredient
T
Phenotype Study
Altered Phenotype

Fig. 15.1 Chemical genetics is used to *validate* a protein for its suitability as a drug *target*. Chemical genetics can be traced back to mutagenesis experiments in genetics. Instead of introducing mutations into the genome, chemical probes are examined for changes in phenotype. Analogous to chemical genomics, the two experimental approaches of forward and reverse chemical genetics exist. Chemical genomics and chemical genetics do not differ in their experimental implementation. They differ only in their goals of identifying and validating the drug target

15.1 Thalidomide and Embryonic Malformation

Thalidomide (alpha-phthalimidoglutarimide) is a chemical probe that was marketed in the 1950s and 1960s under the name thalidomide as a sleeping pill and sedative, especially during pregnancy to treat morning sickness. Thalidomide became sadly famous because it caused malformations of the unborn during pregnancy, which subsequently triggered the thalidomide scandal in the Federal Republic of Germany at the time. Unthinkable today, thalidomide was put on the market without any knowledge of its biochemical mode of action, nor was it tested on pregnant animals before it was approved for marketing. In gross negligence to the latter, it was marketed primarily to pregnant women for morning sickness. Despite this scandal in the 1960s, the mechanism of action of thalidomide remained unknown until 2010.

As shown in Fig. 15.2, the preparation of thalidomide is based on the reaction of phthalic anhydride with the L-amino acid glutamic acid in pyridine. The resulting dicarboxylic acid intermediate is then cyclized to the anhydride in the presence of acetic anhydride. The anhydride is ultimately converted to the imide in reaction with urea. A

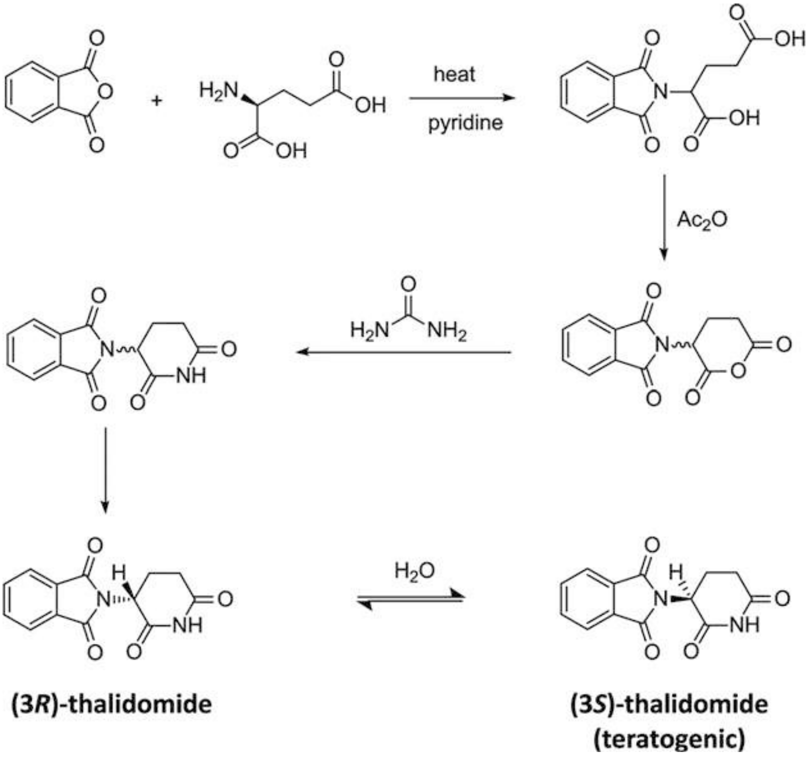


Fig. 15.2 Preparation of thalidomide from phthalic anhydride and L-glutamic acid. A racemic mixture is formed. Even enantiomerically pure thalidomide racemizes in an aqueous solution and the 3S-thalidomide is formed, which has a fruit-damaging effect

racemate of (3S)- and (3R)-thalidomide is obtained. Assuming the structural similarity of the reactant used, the neurotransmitter glutamic acid, a neurological use was suspected and thalidomide was first tested on epilepsy patients. However, thalidomide showed no effect. However, subjects reported sleep-inducing and relaxing properties of thalidomide. These bold assumptions and clinical observations paved the way for thalidomide to enter the healthcare market as a sleep-inducing and sedative agent.

Even long after the withdrawal of thalidomide from the market, the biochemical mechanism that triggered the malformations was unclear. It was not until 2010, more than 50 years after its market introduction, that Japanese scientists published an experiment in which they immobilized thalidomide on a resin sphere analogous to compound-centric chemical proteomics and tested it against a cell lysate. As illustrated in Fig. 15.3, thalidomide binds to the proteins, damaged DNA binding protein 1 (DDB1), and cereblon (CRBN). Binding experiments with isolated cereblon and DDB1 revealed that thalidomide binds to cereblon directly and interacts only with DDB1 via cereblon.

The identification of the E3 ubiquitin-protein ligase cereblon as a target protein of thalidomide allowed us to infer the biochemical pathway of embryonic malformation. Ubiquitin-protein ligases catalyze the transfer of the protein ubiquitin to another protein.

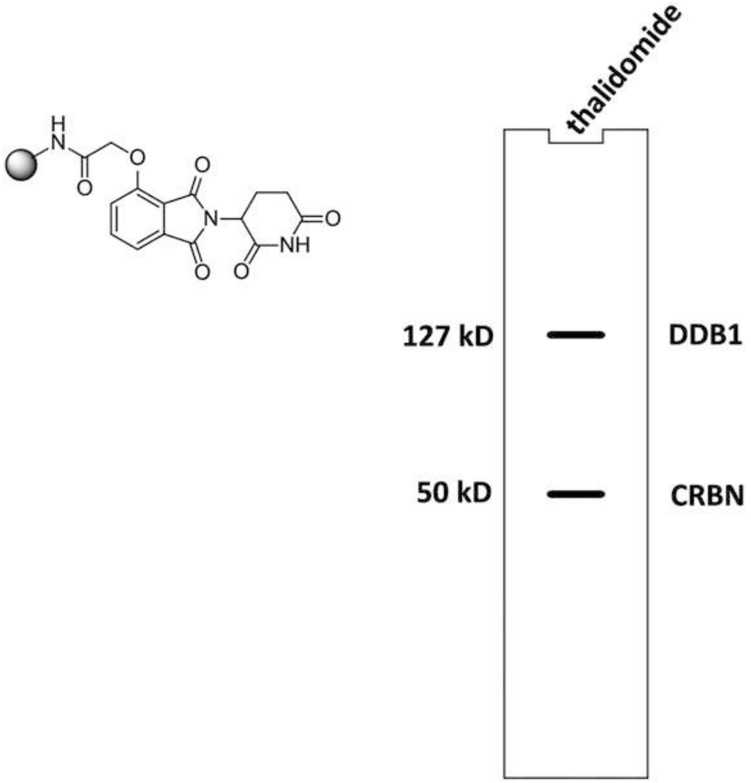


Fig. 15.3 Thalidomide binds to CRBN and DDB1. Thalidomide-binding proteins were isolated from HeLa cell extracts using affinity chromatography

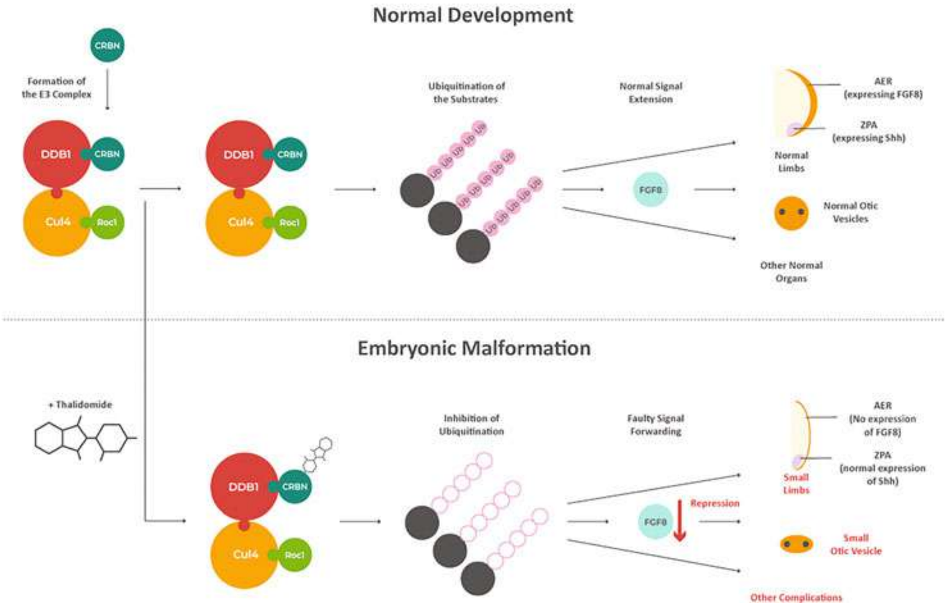


Fig. 15.4 Schematic model for the molecular mechanism of teratogenicity of thalidomide. (**a**) Normally, CRBN functions as a component of the E3 ubiquitin ligase to regulate multiple developmental processes, e.g., limb and otic vesicle formation by ubiquitination of previously unknown substrates. (**b**) Thalidomide binds to CRBN and inhibits the associated E3 function. Aberrant accumulation of the unknown substrate leads to multiple developments of defects, such as small extremities and otic vesicles, in part by downregulating FGF8 expression

The resulting ubiquitin-bound ("ubiquitinated") protein is recognized and degraded by the ubiquitin-proteasome system.

Ubiquitin-protein ligases occur in all cells ("ubiquitous" = everywhere) and ligate ubiquitin to proteins to be degraded in a three-step process with different ubiquitin-protein ligases in each case. In the first step, an E1 ubiquitin-protein ligase catalyzes the binding of the carboxy group (COOH group) of the C-terminal glycine of ubiquitin to a thiol group (SH group) of a specific cysteine on the surface of the E1 ubiquitin-protein ligase by hydrolysis of adenosine triphosphate (ATP). Subsequently, the E1 ubiquitin-protein ligase transfers the ubiquitin to the thiol group of a particular cysteine of the E2 ubiquitin-protein ligase. In the first two steps, the enzymatic transfer of ubiquitin to the enzymes themselves occurs, while in the third step, an E3 ubiquitin-protein ligase catalyzes the transfer of ubiquitin from the E2 ubiquitin-protein ligase to the ε-amino group of a particular lysine on the surface of the protein to be labeled.

If ubiquitination is disrupted, as in the case of thalidomide, the proteins to be degraded remain and defective signal transduction is the result (Fig. 15.4). The defective formation of the limbs in the presence of the thalidomide is primarily due to the reduced expression of the FGF8 gene and thus the absence of the AER protein on the cell surface as well as the lack of formation of the otic or auditory vesicle in the embryo, which ultimately leads to the malformations.

15.2 FK506 and the Ca^{2+}-Calcineurin Signaling Pathway

Probably the best-known example in chemical genetics is FK506 (also known as tacrolimus) and the Ca^{2+}-calcineurin pathway. FK506 is a 23-membered macrolide lactone that was first isolated in 1987 from a Japanese soil sample containing the bacterium *Streptomyces tsukubaensis*. FK506 received special attention when it showed a previously unknown, very potent immunosuppressive effect. In addition to the treatment of autoimmune diseases, such as neurodermatitis, psoriasis, or inflammatory bowel diseases, very strong immunosuppressants play a decisive role in organ transplantations. Only the subsequent drug therapy with FK506 enabled the suppression of organ transplant rejection by the immune system and thus organ transplants at all. Similar to FK506, the structurally related macrolide lactone rapamycin and the cyclopeptide ciclosporin exhibit similarly potent effects as immunosuppressants (Fig. 15.5). Like FK506, rapamycin and ciclosporin are also natural products which, in the case of rapamycin, were isolated from the bacterium

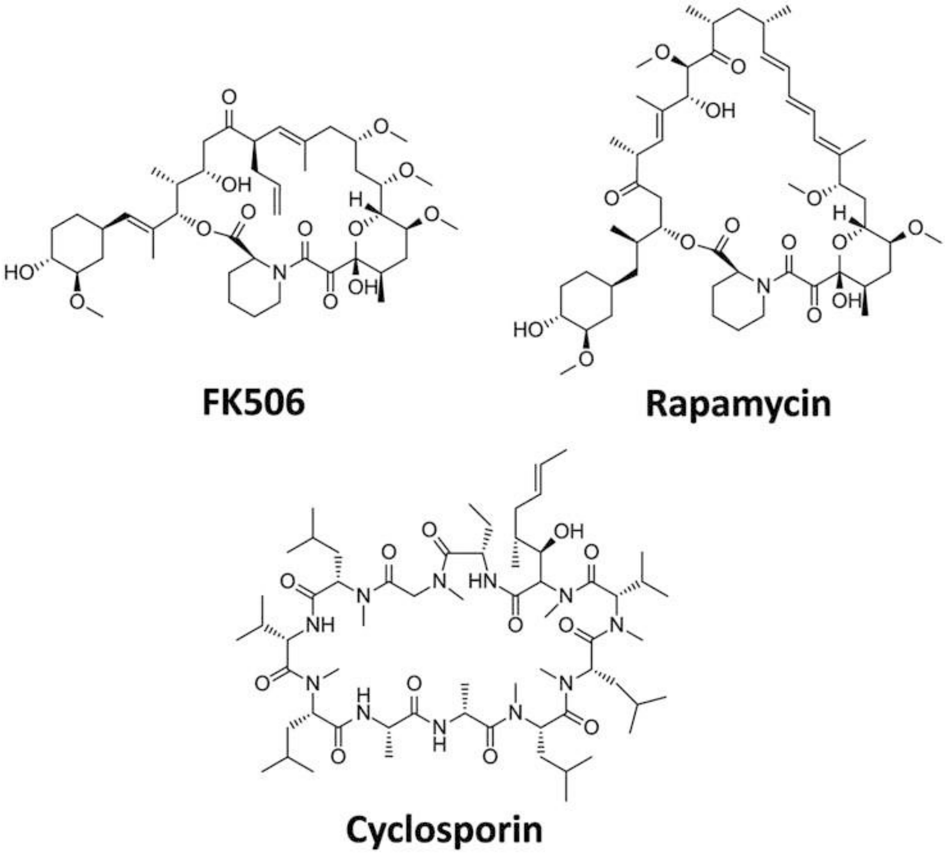


Fig. 15.5 Chemical structures of the immunosuppressive drugs FK506, rapamycin, and ciclosporin

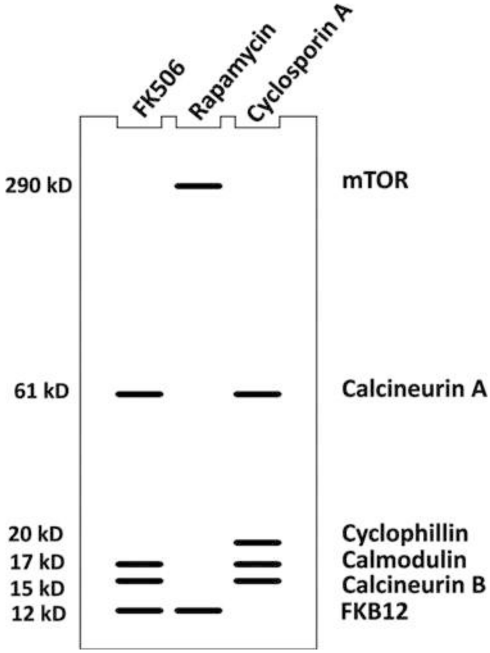


Fig. 15.6 Result of affinity chromatography of the natural products FK506, rapamycin, and ciclosporin. Although all three natural products show immunosuppressive activity, they bind to different proteins. Interestingly, there are overlaps of the binding partners

Streptomyces CC and, in the case of ciclosporin, from the Norwegian sac fungus *Tolypocladium inflatum* and *Cylindrocarpon lucidum.*

Based on the experience with thalidomide from the 1960s, it was necessary to clarify the biochemical mechanism of the three natural substances in order to obtain marketing authorization. As in Sect. 15.1, the three natural products were immobilized and tested in affinity chromatography against a cell extract of immune cells. As shown in Fig. 15.6, the three natural products do not bind the same proteins. However, the binding interactions of the three natural products overlap with the total of six proteins found. FK506 binds to calcineurin A and B as well as to FK-binding protein 12 (FKB12). Rapamycin binds to mTOR and FKB12. Ciclosporin binds similarly to FK506 to calcineurin A and B, calmodulin, but not to FKB12 but cyclophilin. The different binding profiles of the three natural products indicate different mechanisms of modulation of a signaling pathway.

Based on the six binding partners found, further affinity chromatography was used to elucidate the entire Ca^{2+}-calcineurin signaling pathway in T cells. As shown in Fig. 15.7, the T cell is activated via the T cell receptor on its surface. The membrane-bound activated receptor stimulates the release of calcium ions (Ca^{2+}). These bind to calcineurin (CaN), which activates the transcription factor NF-AT. The latter triggers the immune response through the expression of interleukin 2 (IL2). FK506 and ciclosporin do not bind directly to calcineurin. FK506 binds to FKBP12 and ciclosporin binds to cyclophilin (CpN). FK506 and ciclosporin are so-called dimerization inducers. Without the intervention of FK506 and ciclosporin, FKBP12 and cyclophilin do not bind to calcineurin. Only when the two natural

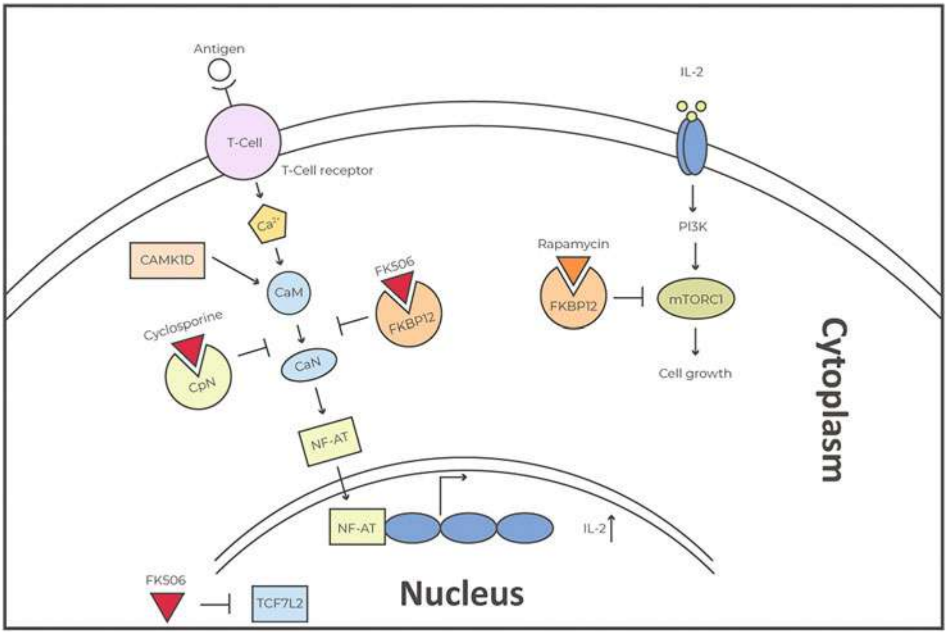


Fig. 15.7 Signaling pathway of the immune response in a T cell. The activated T cell receptor releases Ca^{2+} ions that bind to calcineurin, triggering the binding of the transcription factor NF-AT. As a result, the signaling protein interleukin-2 (IL2) is synthesized and released as part of the immune response. Other T cells bind IL2 and their cell growth is stimulated. In contrast, ciclosporin binds cyclophilin and FK505 binds FKBP12. The binding of the two natural products in both cases causes a conformational change, resulting in the formation of the respective tertiary complex with calcineurin. Activation of the transcription factor NF-AT is absent. Analogously, rapamycin also binds to FKBP12. However, the conformation of FKBP12 changes in such a way that a tertiary complex with mTOR is formed. The absorbed IL2 signal is inhibited

substances bind to them, they change their conformation and can thus bind to calcineurin. As a result, signal transduction of the Ca^{2+}-calcineurin pathway is interrupted. No interleukin-2 is produced. The immune response is absent. However, if rapamycin binds FKBP12, a conformational change of FKBP12 occurs, which allows binding to mTOR. The binding to mTOR subsequently inhibits the cell growth of the T cell, triggered by the interleukin-2 signal of another T cell. The immune response is impaired. Interestingly, FKBP12 can be influenced differently in its conformation by a low-molecular compound to finally bind to different proteins. This process, triggered by a low-molecular compound, is called chemical-induced dimerization (see also Sect. 16.2).

The example of the immunosuppressive natural products FK506, rapamycin, and ciclosporin shows how the mechanisms of the immune response of a T cell could be elucidated from an initially pharmacological question.

Summary

Chemical genetics can be considered analogous to mutagenesis experiments in classical genetics in order to investigate signaling pathways. However, instead of completely eliminating the gene product by mutagenesis, an inhibitor is used that specifically and temporarily disrupts the function of the gene product. The best-known examples of studies in chemical genetics are thalidomide, better known as thalidomide, and FK506, which helped to elucidate entire signaling pathways.

Further Reading

Hardling MW et al (1989) A receptor for the immunosuppressant FK506 is a cis-trans peptidyl-prolyl isomerase. Nature 341:758–760

Ito T et al (2010) Identification of a primary target of thalidomide teratogenicity. Science 327:1345–1350

Liu J et al (1991) Calcineurin is a common target of cyclophilin-cyclosporin A and FKBP-FK506 complexes. Cell 66:807–815

Chemical Biology: Addressing New Drug Targets 16

Despite successes in genome research, the development of new drugs remains a challenge. The central problem is that the majority of genes and their gene products cannot be addressed with the classical tools of drug development—small-molecule ligands and biologics. Small molecule ligands require deep, hydrophobic pockets of their drug targets. Biologics cannot cross the cell membrane and therefore bind only to extracellular drug targets. It is estimated that only 10% of all gene products show the property of a hydrophobic pocket or extracellular expression at any one time (Fig. 16.1). However, finding new drug targets is essential for the development of new therapies.

The systematic screening of drug libraries of small-molecule substances or biologics against all gene products with the aim of identifying new drug targets is the task of chemical genomics. While chemical genetics goes beyond the actual binding to validate the biological effect of the binding and thus its suitability as a drug target, the generic term chemical biology now covers strategies alternative to the classical pharmacological concept of small-molecule substances and biologics that can address drug targets beyond hydrophobic pockets and extracellular expression.

16.1 Protein-Protein Interaction

In general, protein-protein interactions (PPI) are the binding of two or more proteins. Similar to ligand-receptor complexes, the mutual binding of proteins is based on non-covalent interactions, such as Van der Waals forces, hydrogen bonds, electrostatic interactions, and hydrophobic effects of amino acids on the surface of the proteins.

There are approximately 350,000 protein-protein interactions in the cell. Many play a central role in intracellular signal transduction and have been validated as drug targets by mutation studies in cell assays. However, as shown in Fig. 16.2, most PPIs occur over a

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6_16

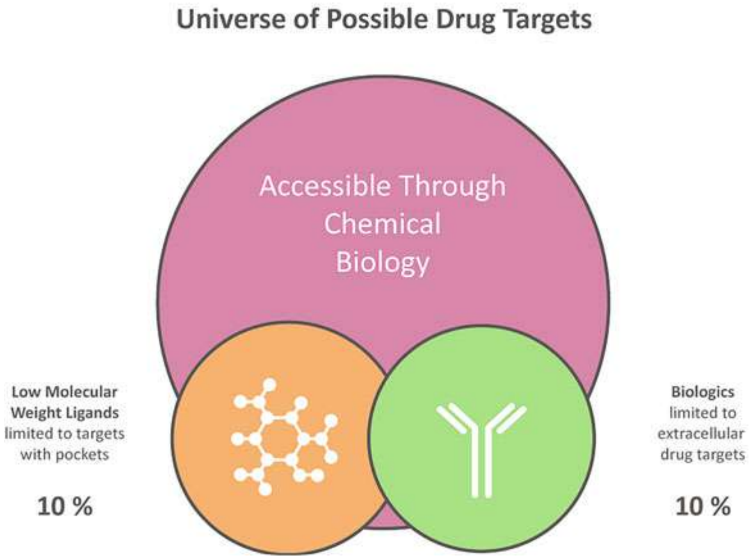

Universe of Possible Drug Targets
Accessible Through Chemical Biology
Low Molecular Weight Ligands limited to targets with pockets
10 %
Biologics limited to extracellular drug targets
10 %

Fig. 16.1 The problem of non-addressable drug targets. Looking at the proteins encoded in the genome, it is noticeable that only 10% of them either have a deep, hydrophobic pocket in which a small-molecule ligand can bind, or the drug target occurs extracellularly so that a biologic can bind to it. It is believed that 80% of all possible drug targets cannot be addressed with small-molecule ligands or biologics. In contrast, chemical biology holds the promise of being able to modulate previously unaddressable proteins through new concepts of action. In pharmacology, *a* drug manipulates *a* drug target, which induces an altered phenotype. The basis of new active concepts in chemical biology are combinations of chemical probes. Proteins that were previously impossible to manipulate are to be modulated by the network

large surface area of the proteins involved. In contrast, receptors, transporters, and especially enzymes form deep, hydrophobic pockets in which low molecular weight substances bind. For this reason, it is very difficult to develop a small molecule inhibitor of a PPI. As explained earlier, only small molecule compounds offer themselves as inhibitors of intracellular drug targets. Although many PPIs are validated drug targets through mutation studies, only one drug that inhibits an intracellular protein-protein interaction has been approved to date: Venetoclax inhibits the protein Bcl-2. Bcl-2 blocks natural cell death (apoptosis). Inhibition of this protein induces apoptosis. Venetoclax is approved for the treatment of blood cancer.

Looking at the chemical structure of Ventoclax, it is clear (Fig. 16.3) what challenges PPIs to pose to drug development. Compared to Lipinsiki's Rule of Five, larger ligands are required, which have to interact with the protein over a larger surface area. On the other hand, the ligands must not become too large, otherwise, they can no longer cross the cell membrane.

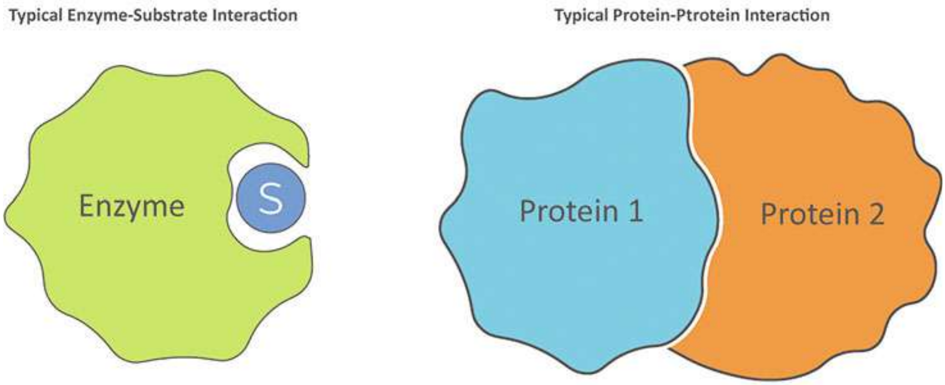

Typical Enzyme-Substrate Interaction
Typical Protein-Ptrotein Interaction
Enzyme
S
Protein 1
Protein 2

Fig. 16.2 Typical enzyme-substrate interaction compared to protein-protein interaction. The mostly low-molecular substrate binds in a deep pocket of the enzyme, whereas a protein-protein interaction occurs over a large surface area of the proteins involved

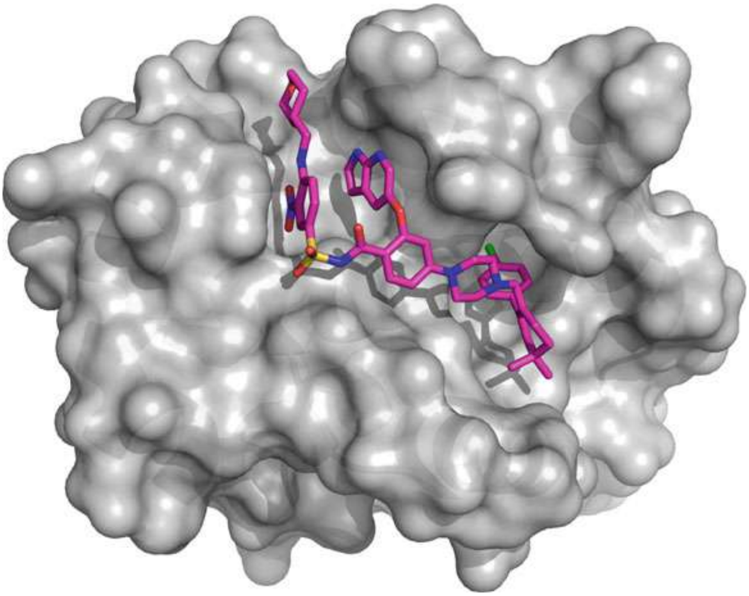


Fig. 16.3 Structure of the only approved protein-protein interaction inhibitor to date, Ventoclax. One can guess the flat binding pocket

Protein-protein interactions currently represent the greatest hope for new drugs, particularly in oncology. However, the development of ligands that can displace the natural binding partner over a large and flat interaction surface while being small enough to still cross the cell membrane is a challenge for medicinal chemistry.

16.2 Chemical-Induced Dimerization (CID)

Chemical-induced dimerization (CID) has already been described in Sect. 16.2. It is a process in which two proteins bind together only in the presence of a specific small molecule.

Chemically induced dimerization was first described for the natural products rapamycin and FK506 (see also Sect. 15.2). The binding of the two proteins occurs only if the dimerization inducer (usually a low molecular weight compound) binds to one of the proteins and changes its conformation so that the protein-protein binding surface is formed and complex formation occurs (Fig. 16.4). As shown in Table 16.1, a number of chemically-induced dimerization are now known. CIDs are used in basic research, for example, to control protein localization or to induce protein activation. In addition, CIDs lend themselves to inhibiting protein-protein interactions. Several companies have, therefore, attempted in the past to manipulate the known CID system of FKBP to bind and inactivate the ras oncogene. The ras oncogene is a protein that plays an important role in many types of cancer but for which no small-molecule inhibitor has yet been found.

Chemically induced dimerization allows inhibition of proteins that do not bind small molecule inhibitors due to their structure without a binding pocket (Fig. 16.5).

16.3 Pharmacological Chaperones

Chaperones are proteins that help other proteins to fold correctly. Pharmacological chaperones are analogous to small-molecule ligands that stabilize mutated misfolded proteins in their active conformation.

Pharmacological chaperones are mostly used for enzymes with a mutation in the protein sequence that leads to misfolding of the enzyme. The misfolded enzyme is normally recognized by the cell's quality control system, retained in the endoplasmic reticulum, and often destroyed or recycled. As a result, the catalytic task of the enzyme is not accomplished. Pharmacological chaperones correct the folding of misfolded proteins by stabilizing the natural conformation through their binding (Fig. 16.6) so that they can pass through the cell's quality control system and be correctly routed. The best-known diseases based on misfolding caused by mutations are cystic fibrosis and Fabry disease from the group of lysosomal storage diseases.

The cause of cystic fibrosis is a malfunction of chloride channels of certain body cells caused by mutation. The cells cannot conduct water into the surrounding tissue by means of osmosis. As a result, the water content of, for example, bronchial secretions, pancreatic secretions, liver secretions, and small intestine secretions is reduced. The secretions become thick and dysfunctions occur in the affected organs. Ivacaftor is a pharmacological chaperone that stabilizes the G551D mutation of the cystic fibrosis transmembrane conductance regulator (CFTR) chloride channel. However, only about 5% of all affected individuals carry this mutation. That is about 3000 patients worldwide.

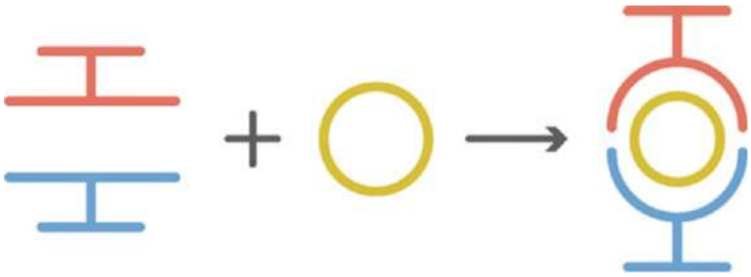


Fig. 16.4 Illustration of the mechanism of chemical-induced dimerization: Two macromolecules can only bind to each other in the presence of the inducer

Table 16.1 Overview of known chemically induced dimerization according to target proteins and dimerization inducers

Target proteins		Dimerization inductor
FKBP	FKBP	FK1012
FKBP	Calcineurin A (CNA)	FK506
FKBP	CyP-Fas	FKCsA
FKBP	FRB domains of the mTOR	Rapamycin
GyrB	GyrB	Coumermycin
GAI	GID1	Gibberellin
Snap tag	HaloTag	HaXS
14-3-3	PMA2	Fusicoccin

Another representative of the pharmacological chaperones is migalastat for the therapy of Fabry disease. Fabry disease belongs to the lysosomal storage diseases. The disease is caused by the deficiency of the enzyme alpha-galactosidase A. As a result, the natural substrate globotriaosylceramide accumulates primarily in the endothelial cells. It is now believed that the accumulation of globotriaosylceramide is not the actual trigger but the product of an alternative metabolic pathway that has not yet been fully elucidated. Similar to ivacaftor, the pharmacological chaperone migalastat stabilizes mutant alpha-galactosidases A. Migalastat (1-deoxygalactonojirimycin: DGJ) is an iminosugar which, as an analog of the terminal galactose of the substrate Gb3, binds to alpha-galactidose A as a reversible inhibitor.

A major advantage of pharmacological chaperones is that, unlike enzyme replacement therapies, they can cross the blood-brain barrier. A disadvantage of pharmacological chaperones is that they are only effective in a small subgroup of patients who form the misfolded proteins. In cystic fibrosis, this is only 5% of patients.

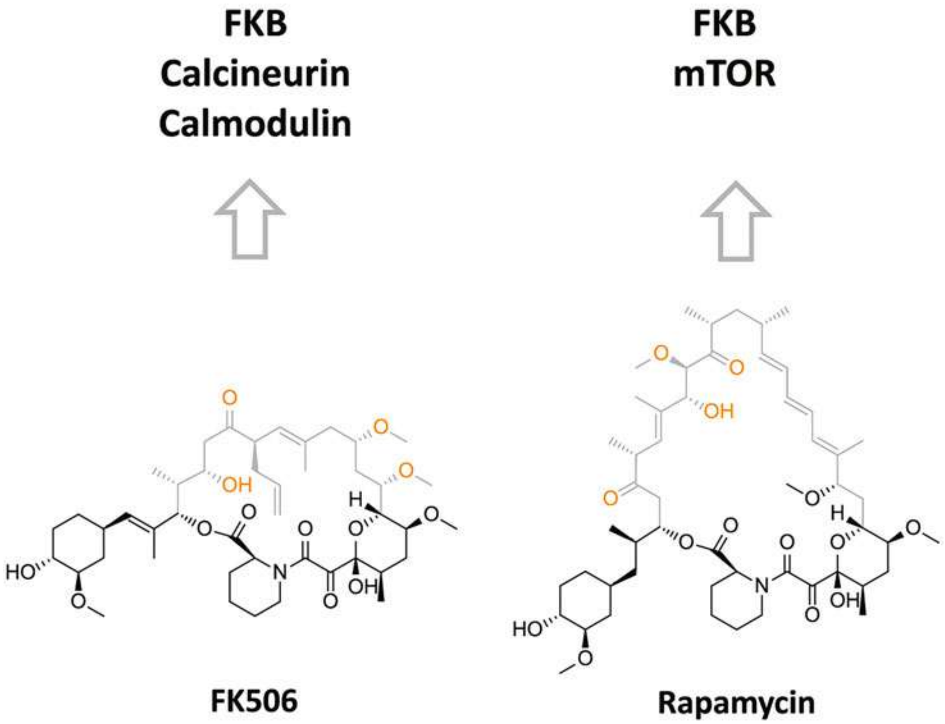


Fig. 16.5 Structures of the two dimerization inducers FK506 and rapamycin have an identical backbone (in black) and a flexible binding motif (in orange). The idea suggests itself chemically modify the flexible binding motif based on the basic structure in order to be able to address other proteins

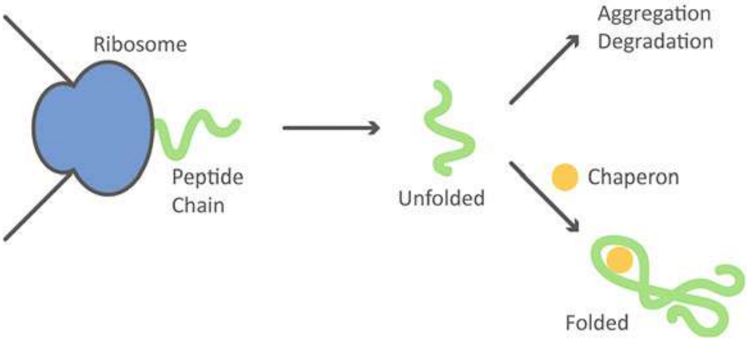


Fig. 16.6 The nascent protein is initially unfolded. A pharmacological chaperone helps to adopt the correct fold. If this is not done, the cell recognizes the misfolding based on the hydrophobic residues on the surface of the protein and degrades it

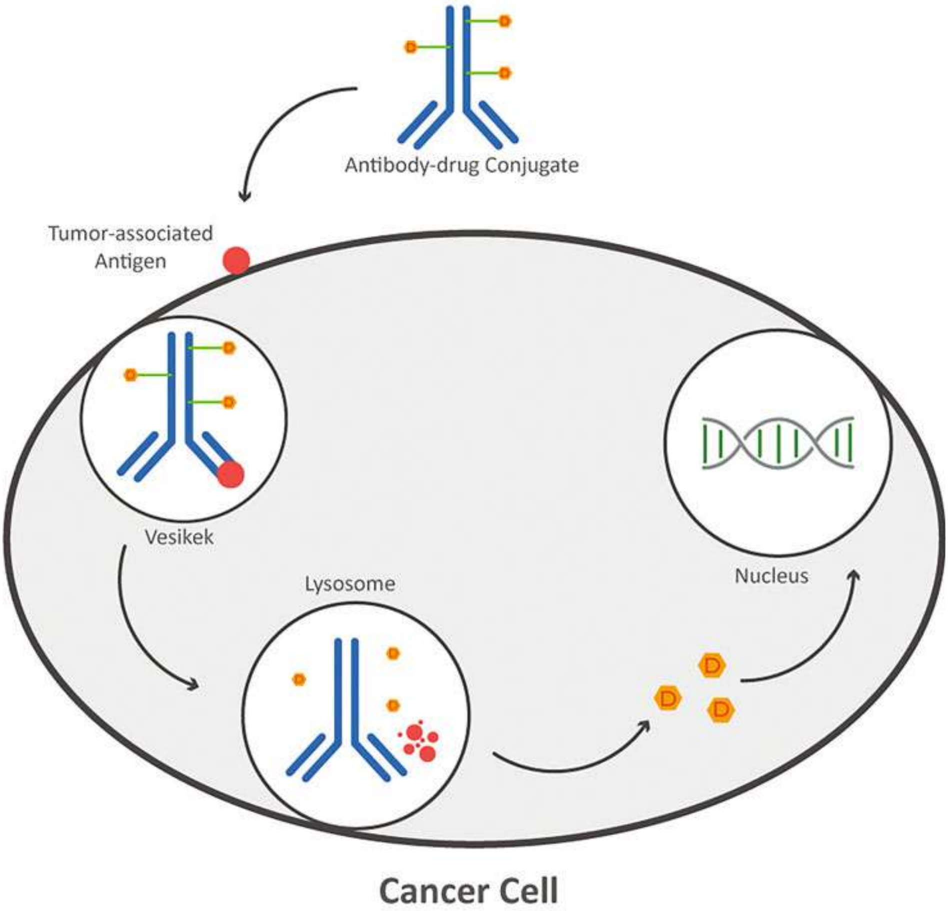


Fig. 16.7 Mechanism of an antibody-drug conjugate. The conjugate binds via its antibody to a tumor-associated antigen on the surface of the cancer cell. Subsequently, the conjugate is internalized. In the lysosome, specific enzymes cleave the linker, the cytostatic drug is released and migrates by diffusion into the cell nucleus

16.4 Antibody-Drug Conjugates

Antibody-drug conjugates (ADCs) are a hybrid class of monoclonal antibodies and small molecule drugs, mostly cytostatics.

In conventional chemotherapy, toxic cytostatic drugs also affect healthy tissue. The aim is to avoid this by coupling to a monoclonal antibody. The antibody-drug conjugates bind in a targeted manner mediated by the antibody only to cancer cells. After binding, the conjugate is internalized and degraded in the lysosome. In this process, lysomal enzymes cleave the linker between the antibody and the cytostatic drug (Fig. 16.7). The low-

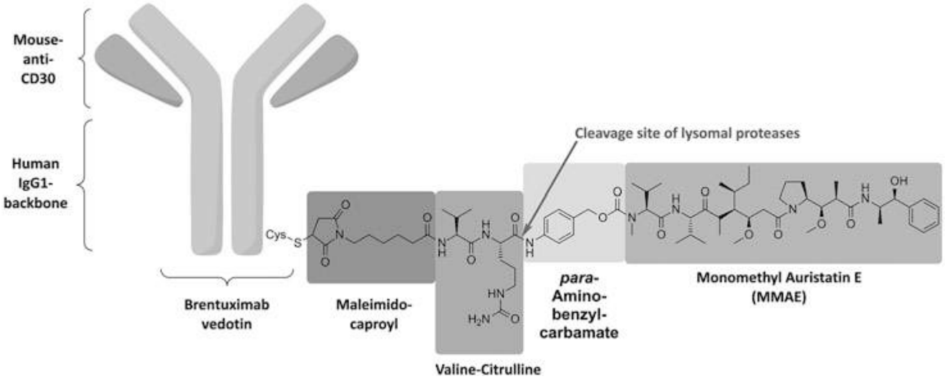

Mouse-anti-CD30
Human IgG1-backbone
Cys
S
Brentuximab vedotin
Maleimido-caproyl
Valine-Citrulline
Cleavage site of lysosomal proteases
para-Amino-benzyl-carbamate
Monomethyl Auristatin E (MMAE)

Fig. 16.8 Structure of an antibody-drug conjugate using the example of brentuximab vedotin

molecular-weight cytostatic is released in the cancer cell and can reach the cell nucleus by diffusion, where it usually binds to DNA and thus interferes with further cell division.

Through the combination of monoclonal antibodies and cytostatics, healthy tissue can be distinguished from diseased tissue. Targeted therapy is possible.

The example of Brentuximab vedotin illustrates the principle of the conjugate (Fig. 16.8): The antibody of the conjugate binds to the specific tumor antigen CD30. The chemotherapeutic agent monomethyl auristatin E (MMAE) is bound to a cysteine of the antibody via a maleimide-valine-citrulline linker. The valine-citrulline motif is recognized by lysosomal proteases, resulting in the release of MMAE.

Antibody-drug conjugates are already being used successfully in the clinic. The five best-known conjugates are listed in Table 16.2.

16.5 Chimeric Antigen Receptor T-Cell Therapy: CAR-T

A consistent derivative of the antibody-drug conjugates is the chimeric antigen receptor T-cell—CAR-T. Similar to the antibody-drug conjugates, the property of a protein (receptor) is used to bind specifically to an antigen of a cancer cell. In this particular case, the receptor protein has been modified in such a way that, on the one hand, it recognizes the antigen of the cancer cell and, on the other hand, it has T-cell activating properties. This is, therefore, referred to as a chimera.

The goal of CAR-T immunotherapy is to modify T-cells to recognize cancer cells in order to target and destroy them more effectively. First, one harvests T-cells from humans. After isolation from a person, these T-cells are genetically modified with a retroviral vector, and more recently with CRISPR/Cas9, to express a specific CAR that programs them to target an antigen found on the surface of tumors (Fig. 16.9). Finally, they are injected into the patient. CAR-T cells can be derived either from T cells in a patient's own blood

Table 16.2 Antibody-drug conjugates approved to date in chronological order of approval

Antibody-drug conjugate	Trade name	Antigen	Cytostatic	Indication
Gemtuzimab ozogamicin	Mylotarg	CD33	Calicheamicin	Acute myeloid leukemia (AML)
Brentuximab vendotin	Adcetris	CD30	Monomethyl auristatin E	Hodgkin's lymphoma/anaplastic large cell lymphoma
Trastuzumab emtansine	Kadcyla	HER2	Mertansin	HER2-positive metastatic breast cancer (mBC) after treatment with trastuzumab
Inotuzumab ozogamicin	Bespona	CD22	Calicheamicin	CD22-positive B-precursor cells of acute lymphoblastic leukemia
Moxetumomab pasudotox	Lumoxiti	CD22	Pseudomonas exotoxin A	Relapsed or refractory hairy cell leukemia
Polatuzumab vedotin-piiq	Polivy	CD79b	Monomethyl auristatin E	Relapsed/refractory diffuse large B-cell lymphoma

(autologous) or from T-cells of another healthy donor (allogeneic). For safety reasons, CAR-T cells carry only specific antigens expressed on a tumor and not on healthy cells.

The Chimeric Antigen Receptor

Crucial to CAR T cell therapy was the development of the chimeric antigen receptor. Chimeric antigen receptors (CARs) combine many facets of normal T-cell activation into a single protein. They link an extracellular antigen recognition domain to an intracellular signaling domain that activates the T-cell when an antigen is bound. CARs consist of four regions: An antigen recognition domain, an extracellular hinge region, a transmembrane domain, and an intracellular T-cell signaling domain (Fig. 16.10).

The antigen recognition domain:

The antigen recognition domain is attached to the outside of the cell in the ectodomain portion of the receptor. This interacts with potential targets and is responsible for the CAR-T cell binding to those cells that express the appropriate target antigen.

The antigen recognition domain is typically derived from the variable regions of a monoclonal antibody linked as *single-chain variable fragments* (scFv). An scFv is a chimeric protein, consisting of immunoglobin light (V_L) and heavy (V_H) chains, linked to a short linker peptide. These V_L and V_H regions are selected in advance based on their binding ability to the target antigen (such as CD19). The linker between the two chains consists of hydrophilic residues with glycine and serine segments for more flexibility and glutamate and lysine segments for more solubility.

In addition to scFvs, non-antibody-based approaches were also used to drive CAR specificity. These have typically exploited ligand-receptor pairs that normally bind to each other. Cytokines, innate immune receptors, TNF receptors, growth factors, and structural proteins have all been successfully used as recognition domains for CAR antigens.

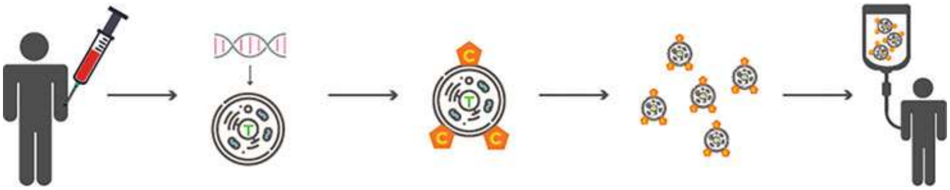
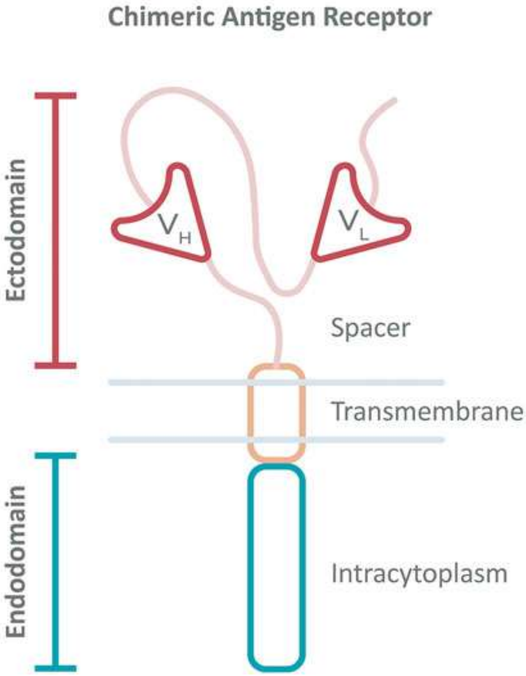


Fig. 16.9 The diagram shows the process of T-cell therapy with chimeric antigen receptors (CAR). This is a method of immunotherapy that is playing an increasingly important role in cancer treatment. The result should be a production of equipped T-cells that can recognize and fight the infected cancer cells in the body. 1. T cells (represented by objects marked with “T”) are removed from the patient’s blood. 2. Then, in a laboratory, the gene that codes for the specific antigen receptors is inserted into the T-cells. 3. This creates the CAR receptors (marked as c) on the surface of the cells. 4. The newly modified T-cells are then further harvested and grown in the laboratory. 5. After a certain time, the engineered T-cells are infused back into the patient

Fig. 16.10 Overview of the structure of the chimeric antigen receptor



Hinge Region

The hinge, also called a hinge or spacer, is a small structural domain located between the antigen recognition region and the outer membrane of the cell. An ideal hinge increases the flexibility of the scFv receptor head and reduces the spatial constraints between the CAR and its target antigen. This promotes antigen binding and synapse formation between CAR

T-cells and target cells. Joint sequences are often based on membrane-proximal regions of other immune molecules, including IgG, CD8, and CD28.

The Transmembrane Domain

The transmembrane domain is a structural component consisting of a hydrophobic alpha helix that spans the cell membrane. It anchors the CAR to the plasma membrane and connects the extracellular hinge and antigen recognition domains to the intracellular signaling region. This domain is essential for the stability of the receptor as a whole. In general, the transmembrane domain is used by most membrane-proximal components of the endodomain but different transmembrane domains lead to different receptor stability. It is known that the CD28 transmembrane domain leads to a highly expressed, stable receptor.

Intracellular T-Cell Signaling Domain

The intracellular T-cell signaling domain is located in the endodomain of the receptor within the cell. After an antigen binds to the external antigen recognition domain, the CAR receptors assemble and send an activation signal. Then, the internal cytoplasmic end of the receptor continues signal transduction within the T-cell.

Normal T-cell activation relies on the phosphorylation of immunoreceptor tyrosine-based activation motifs (ITAMs) present in the cytoplasmic domain of CD3 zeta. To mimic this process, the cytoplasmic domain of CD3-Zeta is commonly used as the major component of the CAR endodomain. Other ITAM-containing domains have also been tried but are not as effective.

In addition to CD3 signaling, T-cells also require co-stimulatory molecules to persist after activation. For this reason, the endodomains of CAR receptors typically also contain one or more chimeric domains of co-stimulatory proteins. Signaling domains from a variety of co-stimulatory molecules have been successfully tested, including CD28, CD27, CD134 (OX40), and CD137 (4-1BB).

Evolution of the Chimeric Antigen Receptor

The first CAR-T-cells were developed as early as the late 1980s. The further development of the constructed CAR receptors has grown over time and is referred to as first, second, third, or fourth-generation CAR, depending on their composition.

First-generation CARs consist of an extracellular binding domain, a hinge region, a transmembrane domain, and one or more intracellular signaling domains (Fig. 16.11). The extracellular binding domain contains a single-chain variable fragment (scFv) derived from tumor antigen-reactive antibodies and usually has high specificity for a given tumor antigen. All CARs contain the CD3 chain domain as an intracellular signaling domain, which is the primary transmitter of T-cell activation signals.

Second-generation CARs add a co-stimulatory domain, such as CD28 or 4-1BB. The involvement of these intracellular signaling domains enhances T-cell proliferation, cytokine secretion, apoptosis resistance, and in vivo persistence.

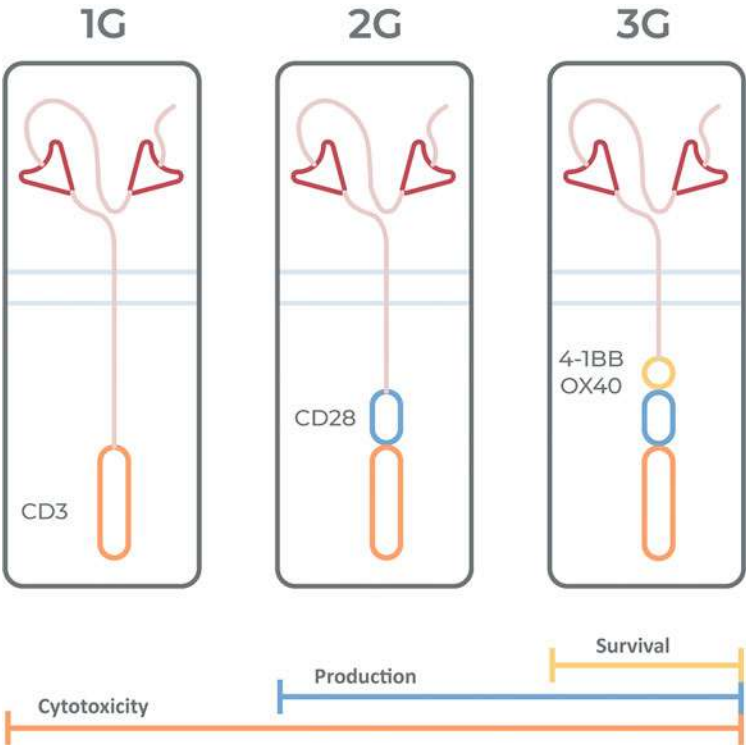


Fig. 16.11 Representation of first-, second-, and third-generation chimeric antigen receptors with the scFv segments in red and the various TCR signaling components in orange, blue, and yellow

Third-generation CARs combine multiple co-stimulatory domains, such as CD28-41BB or CD28-OX40 to enhance T-cell activity. Preclinical data show that third-generation CARs have improved effector functions and in vivo persistence compared to second-generation CARs.

Fourth-generation CARs (also referred to as TRUCKs or armored CARs) add additional factors that enhance T-cell expansion, persistence, and antitumor activity. This may include cytokines, such as IL-2, IL-5, IL-12, and co-stimulatory ligands.

16.6 Hydrophobic Tagging for Targeted Protein Degradation: Halo Tag

A more advanced concept of protein degradation is hydrophobic tagging using the halo tag. While chemical-induced dimerization requires the presence of a hydrophobic binding pocket to address the target protein, hydrophobic tagging for targeted protein degradation can be performed without a binding pocket of the target protein. However, this requires the insertion of a halo-tag fusion protein using gene therapy or genome editing.

The Halo Tag

Halo tag is originally a 297 amino acid long bacterial haloalkane hydrolase with a genetically modified active site that covalently binds the reactive chloroalkane linker with high affinity by means of a nucleophilic substitution reaction 2 (S_N2). The reaction is irreversible under physiological conditions.

The chloroalkane binds specifically in the substrate pocket of haloalkane hydrolase and is fixed by tryptophan and an asparagine residue, resulting in an S_N2 reaction by the activated aspartate (Fig. 16.12). In the wild-type enzyme, the ester is cleaved again by activated water. In contrast, the mutant halo tag cannot hydrolyze the ester. A covalent ester bond remains between the substrate and the halo tag.

Halo tag fusion proteins can be introduced using standard techniques for recombinant protein expression. In addition, several commercial vectors are available that require only the insertion of a gene of interest. Because bacterial dehalogenases are relatively small and the reactions described above are foreign to mammalian cells, there is no interference from endogenous mammalian metabolic reactions. Once the fusion protein has been expressed, a haloalkane can be added which covalently binds to the fusion protein.

Using a halo tag fusion protein, the target protein can be systematically degraded after the addition of a small hydrophobic molecule. This process is called hydrophobic tagging for targeted protein degradation.

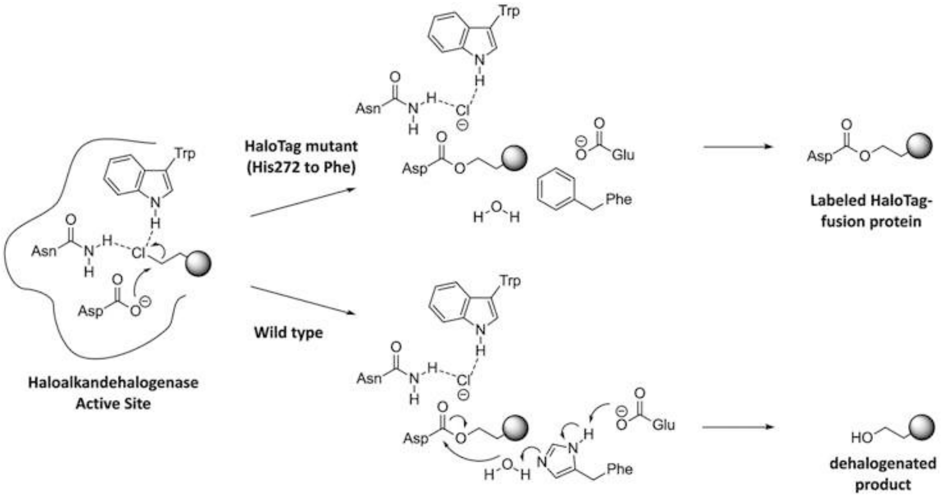


Fig. 16.12 (**a**) Mechanism of haloalkane halogenase: With the mutation of His272 to phenylalanine resulting in the formation of a covalent ester bond between the enzyme and the alkane. (**b**) The catalytic mechanism of the wild type for the dehalogenation of the chloroalkane

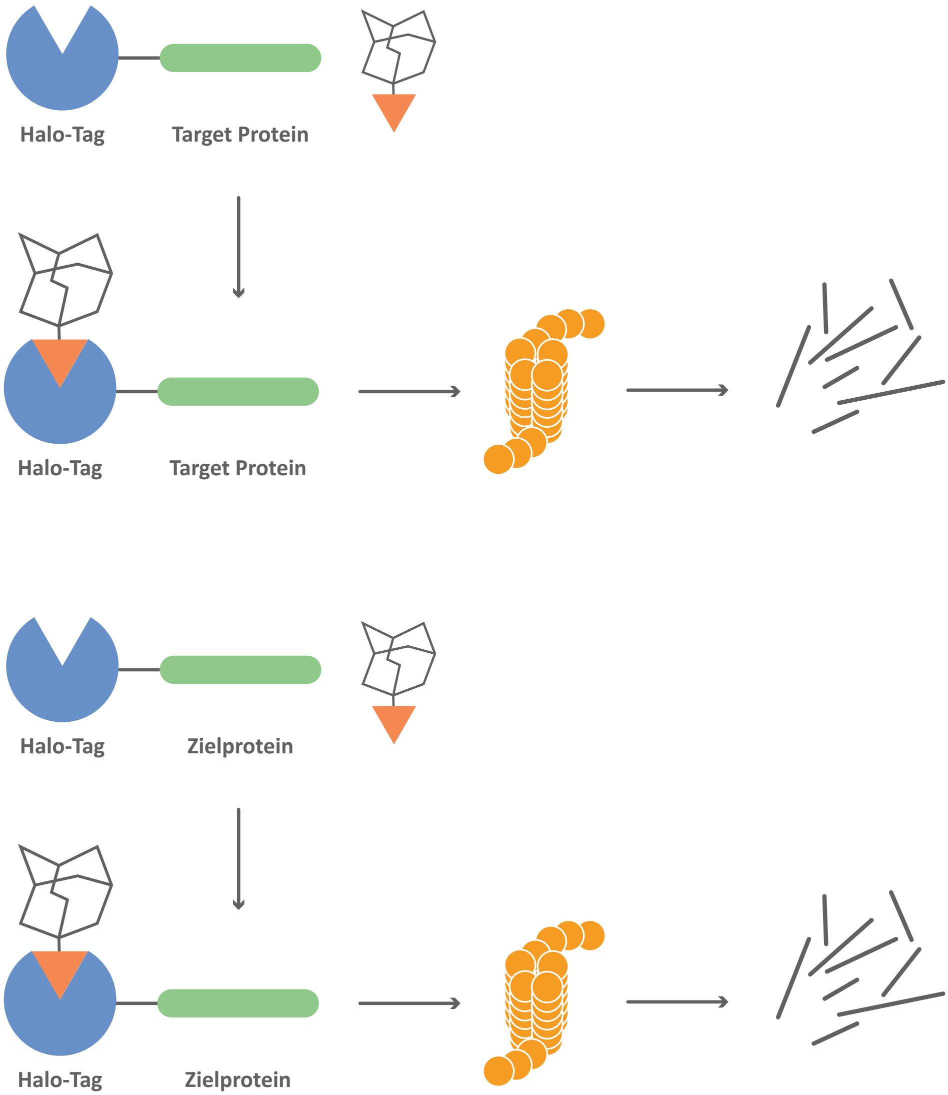


Fig. 16.13 Schematic representation of HyT13-mediated degradation of a fusion protein, consisting of the target protein and the halo tag, by the proteasome

As shown in Fig. 16.13, the target protein is introduced as a fusion protein by gene therapy or genome editing. Only by adding a chloroalkane with a hydrophobic tag, the fusion protein is tagged for degradation by the proteasome. This strategy has several advantages over existing technologies, such as gene therapy, CRISPR-Cas, chemically-induced dimerization, or PROTAC. In contrast to the former two methods, the relevant protein is expressed normally and only when required is specifically released by the addition of a hydrophobic tag for protein degradation. It has been shown that in the absence of the degradation signal, the fusion protein is stable and can perform its native functions. The concept of functional modulation by drugs has been established for this case. However, many proteins do not bind small-molecule ligands or antibodies. For this case of temporary targeted elimination of a protein, the method offers enormous potential.

In summary, hydrophobic tagging for protein degradation is an interesting approach for target proteins that a) do not themselves present a binding surface for a small molecule ligand or antibody and b) present such an essential function that they cannot be completely overwritten using genome editing but can only be temporarily or tissue-specifically deactivated. The procedure has so far only been tested in cells and some small animal models. For an application in humans, the challenge is to find effective vectors for the introduction of the fusion protein as well as the question of which target protein should be addressed.

16.7 Proteolysis Targeting Chimera: PROTAC

Although high-affinity ligands are known for some target proteins, cell experiments have already shown that binding per se does not trigger any biochemical effect. The binding of a ligand at a protein site does not necessarily imply a change in function.

Craig Crews took up this problem at the turn of the millennium. He developed the concept of proteolysis targeting chimera (PROTAC). Ultimately, he synthetically combined two low-molecular ligands. One ligand binds to the target protein. The other ligand binds to the E3 ligase of ubiquitin-mediated protein degradation. This type of ligand is called a chimera, in reference to Greek mythology. As shown in Fig. 16.14, the chimera binds the target protein on the one hand and the E3 ligase on the other. In the process, a tertiary complex is formed. The E3 ligase recruits the E2 ligase, which in turn ubiquitinates the target protein. The ubiquitinated protein is recognized by the cell's protein degradation system and eventually degraded in the proteasome. Meanwhile, the concept has been applied to several E3 ligases, such as pVHL, MDM2, beta-TrCP1, cereblon, or c-IAP1. Currently, a chimera is already in clinical phase II.

PROTAC can be used to inhibit proteins for which high-affinity ligands exist but which have no biochemical effect. This method can be used to inhibit a large number of previously unaddressable drug targets. The space of addressable drug targets increases enormously by this method.

16.8 Old Drugs: New Indications

Repositioning, i.e. finding a new therapeutic application for a known drug, is not new. This is already the case for a total of 14 drugs (Table 16.3). To date, repositioning has taken place in three, rather random ways:

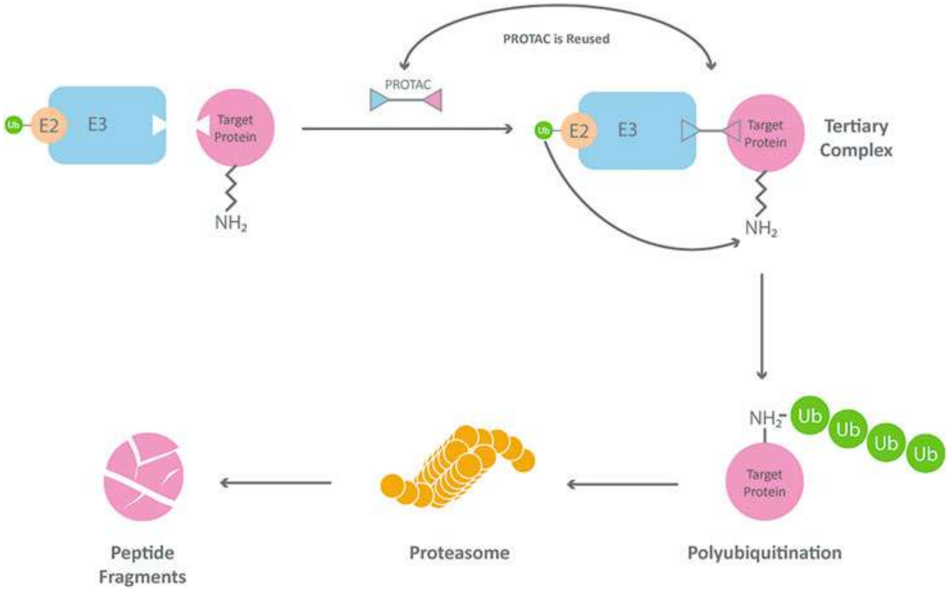


Fig. 16.14 Mechanism of targeted proteolysis by a chimera. The proteolysis targeting chimera (PROTAC) binds to the target molecule and E3 ligase, causing the target protein to be ubiquitinated and degraded. The great advantage of this method is the inhibition of proteins for which low-molecular ligands are known but which cannot inhibit the protein in its function

- in vitro screening of known compounds against new drug targets,
- pharmacological analysis,
- retrospective clinical analysis.

In Vitro Screening of Known Drugs

For newly discovered drug targets, drug libraries with more than one million members are screened in the search for potential drugs. These compound libraries often contain compounds that have already been approved for other indications. The chance of finding a new drug target and thus a new indication for a known drug in this way is extremely small. This is why zidovudine, originally developed in oncology and repositioned for HIV/AIDS, is only one example.

Pharmacological Analysis

While in vitro screening is used to find a new drug target for a known drug, the pharmacological analysis focuses on the known drug target and its clinical effect. Here, one tries to link the drug target or its function in a signaling pathway with new biochemical findings of diseases other than those known so far. The best example of the repositioning of a drug is thalidomide. Originally developed as a treatment for morning sickness in pregnancy, it

Table 16.3 Overview of all drugs successfully repositioned to date

Drug	Original indication	New indication	Approval	Repurposing process	Note
Zidovudine	Oncology	HIV/AIDS	1987	In vitro screening of composite libraries	Zidovudine was the first anti-HIV drug to be approved by the FDA
Minoxidil	Hypertension	Hair loss	1988	Retrospective clinical analysis (identification of hair growth found to be detrimental)	Global sales for minoxidil amounted to US$860 million in 2016
Sildenafil	Angina	Erectile dysfunction	1998	Retrospective clinical analysis	Marketed as Viagra, sildenafil became the leading erectile dysfunction product, with worldwide sales in 2012 of US$ 2.05 billion
Thalidomide	Nausea	Leprosy/ multiple myeloma	1998/ 2006	Off-label use and pharmacological analysis	Thalidomide derivatives show substantial clinical success in multiple myeloma
Celecoxib	Pain/inflammation	Adenomatous polyps	2000	Pharmacological analysis	Total Celebrex sales at the end of 2014 were US$ 2.69 billion
Atomoxetine	Parkinson's disease	ADHD	2002	Pharmacological analysis	Strattera reported global sales of US$855 million in 2016
Duloxetine	Depression	Incontinence	2004	Pharmacological analysis	Duloxetine is only approved in the EU for incontinence. In the USA, the application was withdrawn. There, duloxetine is only approved for the treatment of depression and chronic pain
Rituximab	Oncology	Rheumatoid arthritis	2006	Retrospective clinical analysis (remission of coexisting rheumatoid arthritis in patients with non-Hodgkin's lymphoma treated with rituximab)	Global sales of rituximab exceeded US$ 7 billion in 2015

(continued)

Table 16.3 (continued)

Drug	Original indication	New indication	Approval	Repurposing process	Note
Raloxifene	Osteoporosis	Breast cancer	2007	Retrospective clinical analysis	Approved by the FDA in the USA for invasive breast cancer. Global sales were US$ 237 million in 2015
Fingolimod	Immunosuppressant	MS	2010	Pharmacological and structural analysis	First oral therapy approved for MS. Worldwide sales of fingolimod Reached US$ 3.1 billion in 2017
Dapoxetine	Depression	Premature ejaculation	2012	Pharmacological analysis	Approved in the UK and a number of European countries; to be approved in the USA
Topiramate	Epilepsy	Obesity	2012	Pharmacological analysis	Topiramate is marketed in combination with phentermine
Ketoconazole	Fungal infection	Cushing’s syndrome	2014	Pharmacological analysis	Approved by the EMA for Cushing’s syndrome in adults and adolescents over the age of 12 years
Aspirin	Analgesia	Colorectal cancer	2015	Retrospective clinical and pharmacological analysis	US Preventive Services Task Force published recommendations in September 2015 on the use of aspirin to prevent cardiovascular disease and colorectal cancer

caused a scandal in the 1960s because thalidomide, marketed under the name Contergan, caused malformations of the fetus during pregnancy. This property of developmental inhibition leads to its use in leprosy and multiple myeloma.

Retrospective Clinical Analysis

Retrospective clinical analysis is nothing more than the incidental observation of a new, previously unknown clinical effect of a drug. The most famous example is the drug sildenafil, which was initially developed for the treatment of angina. It was only after several male study participants were asked to continue taking sildenafil beyond the study that it was realized that sildenafil could treat erectile dysfunction.

In contrast, our approach allows the targeted search for new therapeutic applications of known drugs based on the genetic linkage of the drug target (drug target linkage) with a validated analysis technology for wide data and is not dependent on random observations.

Summary

The central problem in drug development today is that for many drug targets identified by genome research, no active agent can be found in the form of a small molecule or biologic. Therefore, approaches are being sought in drug development that allows previously unaddressable drug targets to nevertheless be manipulated with chemical probes. In the pharmaceutical industry, these strategies are now referred to as chemical biology. Promising strategies beyond small molecule drugs and biologics are protein-protein interactions, chemical-induced dimerization, pharmacological chaperones, antibody-drug conjugates, chimeric antigen receptor therapies (CAR-T), hydrophobic tagging for targeted protein degradation, PROTAC, or the discovery of new therapeutic applications of known drugs (repositioning). A major goal of chemical biology is to pursue these approaches further and to develop new strategies.

Further Reading

Beck A et al (2017) Strategies and challenges for the next generation of antibody-drug candidates. Nat Rev Drug Discov 16:315–337

Convertino M, Das J, Dokholyan NV (2016) Pharmacological chaperones: design and development of new therapeutic strategies for the treatment of conformational diseases. ACS Chem Biol 11:1471–1489

Fegan A et al (2010) Chemically controlled protein assembly: techniques and applications. Chem Rev 110:3315–3336

June CH, Sadelain M (2018) Chimeric antigen receptor therapy. N Engl J Med 379:64–73

Neklesa TK et al (2011) Small-molecule hydrophobic tagging-induced degradation of HaloTag fusion proteins. Nat Chem Biol 7:538–543

Pushpakom S et al (2019) Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 18:41–58

Sakamoto KM et al (2001) Protacs: chimeric molecules that target proteins to the Skp1-cullin-F box complex for ubiquitination and degradation. Proc Natl Acad Sci U S A 98:8554–8559

Index

M. F. Schmidt, *Chemical Biology*,
https://doi.org/10.1007/978-3-662-64412-6