The Biological Tools of Modern Chemistry

Béla Gyurcsik

University of Szeged

2020

This teaching material has been made at the University of Szeged and supported by the European Union. Project identity number: EFOP-3.4.3-16-2016-00014

Preface

What is the meaning of the modern chemistry, or rather the modern chemist? In view of the author a description like this could be applied for all the experimental sciences. First of all, the modern researcher recognises a scientific/practical problem. Let's take the development of a drug molecule against cancer as an example. To solve this problem, we have to understand that it shall be approached from various aspects, and in fact it is an interdisciplinary task. We need to apply all the necessary tools (knowledge, instruments) of the chemistry, but also of other scientific fields. In the following is a list of few aspects of this research without an endeavour of completeness. Informatics helps the design of the drug molecule taking into account the many factors by various programs. Biochemistry provides information on the pathway of the drug action and metabolism. Biophysics and bioanalytical experiments tell about the structure and speciation of the drug molecule under various conditions – e.g. upon interaction with physiological solution, membranes, receptors and target biomolecules. Mathematics contributes in validation of the experimental results through statistical analysis of the data. It is also necessary to think about the practical applicability, economical aspects etc. including specialists in medicine and business and many more. Chemists will synthesize and characterize the drug molecule and encapsulate it to reach the biological target. However, we can easily recognize that without the know-how listed above this will not be a successful project. Thus, we need to collaborate with researchers from different fields.

As a chemist, the author has frequently faced difficulty to discuss with biologists, physicists, medical doctors, etc, and recognized that it is very important to precisely understand each other's scientific language. An adventurous postdoctoral research trip to a famous Japanese university, Tokyo Institute of Technology, within the frame of a UNESCO fellowship, provided a chance to the author to get insight into the life of a virology laboratory, extensively applying molecular biology tools in their experiments. We have been mutually teaching chemistry and biology to each other with the Japanese members of the laboratory. The acquired knowledge made possible to synthesize, purify and characterize biological macromolecules, such as proteins and DNA, as well as, to carry out their targeted modification and investigate the effect of the changes on the structure, interaction and function. This experience proved to be fruitful in the establishment of new bioinorganic chemistry approach at University Szeged, namely the study of biomolecules and their derivatives themselves, instead of classical modelling studies.

Recognizing the usefulness of the interdisciplinary outlook, the author decided to invent a new course for the non-biology students to assist the development of their collaborative problem solving approach. Students learn about basics of molecular biology tools that can be applied in advanced chemical and biochemical research. The goal of this electronic book is to provide written material in support of the university course under the same name. The course is primarily suggested for chemistry and info-bionics master and doctoral students. Nevertheless, its biological analytical part serves as a guide to the students in biology, molecular bionics and pharmaceutics through the instrumental methods of the biomolecule identification, purification and analysis.

Contents

General learning outcomes

This book is intended to help students reaching the following general learning outcomes:

Knowledge

The students define the possibility of the overlap of biology and chemistry, list the appropriate borderline research fields.

The students know the biological tools suitable to solve the given complex interdisciplinary problem.

The students summarize the newly introduced/ acquainted methods suitable to study the properties of biological macromolecules.

The students denominate modern procedures, by the help of which macromolecules of biological importance (DNA, RNA, protein) can be synthesized.

The students list the opportunities of applications of the tools of molecular biology in the field of chemistry research.

The students understand the overlap of the various scientific disciplines.

The students explain the effect of molecular biology on the development of modern chemical research.

The students know the methods of examination used in molecular biology.

The students are aware of the background and application opportunities of various microscopic methods.

The students know the theory and practice of the molecular biology tools for studying the structure of protein molecules.

Skills

The students analyse the possibilities of the collaboration with researchers representing various scientific disciplines.

The students communicate with researchers representing biology and other scientific disciplines to solve a complex interdisciplinary problem.

The students evaluate the capability of biological and chemical procedures in the separation and investigation of DNA or protein molecules.

The students carry out the targeted modification of a protein molecule based on its genetic code in theory and practice.

The students formulate their interdisciplinary projects in bioinorganic chemistry, drug design, understanding biological processes, the mechanism of drug action, etc. - all projects involving macromolecules such as proteins and DNA.

The students efficiently discuss with collaborators from biology, pharmacology and medicine field of sciences.

The students select the appropriate molecular biology tools that can be applied in advanced chemical and biochemical research to solve the given complex problem.

The students select the appropriate procedure for the purpose of the detection, identification and purification of DNA or protein molecules.

The students explain the role of the metal ions in biological systems including examples for the role of the free metal ions, metalloproteins and metalloenzymes.

The students select the appropriate methods to investigate the amino acid composition, amino acid sequence.

The students distinguish the various levels of the protein structure, list methods by means of which these structural organizations can be studied.

The students estimate the secondary structure of the protein molecules.

The students demonstrate by examples the possibilities of the modification of biological macromolecules (DNA and protein), including point mutations, deletions or insertions and fusions.

The students explain the steps of the design of biological macromolecules.

Attitude

The students pay attention to the precise application of the chemistry and molecular biology terminology.

The students make effort to apply the interdisciplinary approach in their study and research.

The students help the colleagues from the biology field in understanding the methods of chemical approach to a biological experiment.

The students are motivated to acquire new information on diverse scientific fields.

The students are critical during the evaluation of the literature.

Responsibility and autonomy

The students collaborate with colleagues from biology or various research areas upon recognizing an interdisciplinary research problem.

The students independently apply the proper terminology of chemistry and molecular biology, and explains the terminology to either chemistry or biology orientated colleagues.

Facing an interdisciplinary problem, the students independently develop their knowledge on the borderline scientific areas.

The students understand that the complex research project has to be conducted in collaboration with researchers from other scientific disciplines.

The students can critically evaluate the results of the complex experiments including high number of the degree of freedom.

1. Introduction

At the beginning of the scientific research it was common the observe the universe as a whole and to collect knowledge in every aspect of the observed objects. The polyhistors intended to understand and explain natural phenomena based on their knowledge regardless of the fields of sciences evolved since then. The exponential development of science has been realized, however, in extreme differentiation of various fields. Nowadays students and researchers experience difficulties in understanding each others' presentations, since the research topics are extremely specific to one subject or to one new experimental technique. A biologist colleague of the author told already several years ago that his life-long research topic will be the study of the digestive disorders of the shrimps living in the deep oceans. Indeed, we can meet with such focused projects in an uncountable amount of research articles. These pieces of isolated works may contribute to the understanding of the nature, but they also may cause difficulty and confusion, if they diverge far from the coherency of the multiple processes in our environment.

It might be even worse if the article contains mistakes, misleading information and/or the experiments are described in a way that they are not reproducible. These phenomena are unfortunately not rare. The research is becoming business in an increasing extent. It is natural that the researchers would like to benefit from their discoveries helping to invent new technologies, environmental protection, drugs for diseases leading to better life of people.

However, the introduction of money directly in the publication process lead to controversy concerning the scientific level of the articles appeared in various journals. Also the researchers are pressed to publish more and more since often the indicators such as the number of the published papers, or their citations are takes as the measure of the value of their scientific research. Based on these numbers they are advanced or awarded grants etc. Because of the high number of manuscripts, it is difficult to select proper researchers for the evaluation process being experts on the specific field. Therefore, increasing amount of incorrect information will necessarily appear in the scientific literature. We should think about the request of Max Delbrück, a Nobel-prize holder in his letter to the wife of his friend and colleague, Seymour Benzer – as it is described in the book Csillagórák a tudományban by Venetianer Pál. Citing this part of the letter: "Dear Dotty, please tell Seymour to stop writing so many papers. If I gave them the attention his papers used to deserve, they would take all my time. If he must continue, tell him to do what Erns Mayr asked his mother to do in her long daily letter, namely, underline what is important."

It is also worth mentioning a determinative episode from Delbrück's carrier taken from his biography at the Nobel Prize website for improving the future opportunity of the reader. He has spent three postdoctoral years (1929-1932) abroad, in England, Switzerland, and Denmark. The stay in England, with its immersion into a new language and a new culture, had a vast effect on widening his outlook on life. In Switzerland and Denmark the associations with Wolfgang Pauli and Niels Bohr shaped his attitude toward the pursuit of truth in science.

Apart from the above discrepancies, the high amount of the published information (even restricted to the scientific literature) makes it impossible to

collect all this in a single mind, thus the era of polyhistors is over. Nevertheless, the nature can not take our limitations into account: the problems are complex, as it is schematically shown in **Fig. 1.**

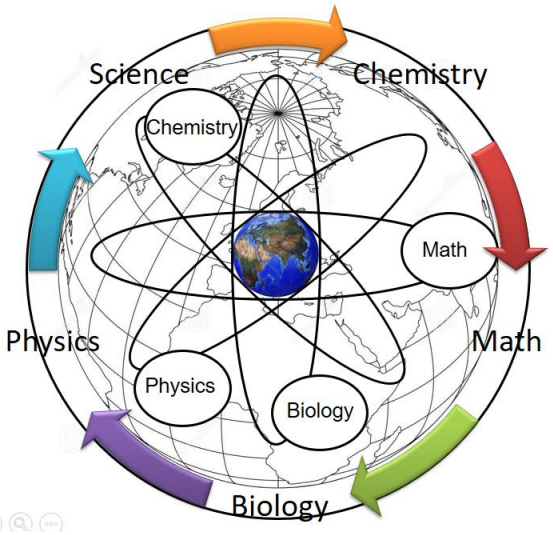


Figure 1. Schematic representation of the complexity of the natural systems, and the need for the collaboration of the researchers for various research fields for correct problem solving.

Therefore, to keep the life going on on earth, we need to collect the scientific knowledge. It is not enough e.g. to develop a more economic technology, but the multiple consequences of its practical application on the environment should be strictly verified. We need to solve the problem, how to gather together enough and reliable information to be able to come to a right decision. Informatics can be an invaluable tool in collecting and analysing the data, but the knowledge to develop such methods and to correctly understand this can only arise from the collaboration of the researchers from diverse scientific fields.

2. The overlap of the chemical and biological sciences

Students who study this chapter will acquire the following specified learning outcomes:

Knowledge

The students define the possibility of the overlap of biology and chemistry, list the appropriate borderline research fields.

The students list the opportunities of applications of the tools of molecular biology in the field of chemistry research.

The students explain the effect of molecular biology on the development of modern chemical research.

Skills

The students formulate their interdisciplinary projects in bioinorganic chemistry, drug design, understanding biological processes, the mechanism of drug action, etc. - all projects involving macromolecules such as proteins and DNA.

The students communicate with researchers representing biology and other scientific disciplines to solve a complex interdisciplinary problem.

The students analyse the possibilities of the collaboration with researchers representing various scientific disciplines.

Attitude

The students make effort to apply the interdisciplinary approach in their study and research.

The students help the colleagues from the biology field in understanding the methods of chemical approach to a biological experiment.

The students are motivated to acquire new information on diverse scientific fields.

The students pay attention to the importance of interdisciplinary research.

Responsibility and autonomy

Facing an interdisciplinary problem, the students independently develop their knowledge on the borderline scientific areas.

The students understand that the complex research project has to be conducted in collaboration with researchers from other scientific disciplines.

With the development of the science including knowledge and research infrastructure an increasing demand arose for collaboration among various disciplines to solve a specific problem. As the result of this evolutionary process, the interdisciplinary outlook has become indispensable in scientific work. Nowadays a large number of procedures, which were inaccessible just several years ago, are routinely carried out. This provides a unique chance to extend our experimental repertoire and to promote our research into the domain of modern science. Day by day we witness the appearance of new names for various interdisciplinary fields followed by the new editions of scientific journals and books.

Bioinorganic chemistry is an interdisciplinary science emerged to connect the knowledge accumulated in inorganic chemistry and biology. The knowledge about the presence of the inorganic compounds, with main emphasis on the metal ions has been extended by the development of new analytical methods. Several metal ions have been found to be an integrated part of the biological fluids, humors, cells, and biomolecules. At the same time few metal ions acquired from the environment may exert beneficial or toxic effect. It became the important to keep tabs on the modes of their intake, distribution and function of all these metal ions to understand their biological role. An inorganic chemist has the knowledge about the physicochemical properties of the elements and their compounds, which can be used to explain the strength of the interactions, the local structure of the metal ion surroundings, and the reactivity of the metal ions and/or the molecules interaction with them.

Metal ions may have been involved in the formation of the first biomolecules through complex formation processes, which arranged the ligand molecules in a favourable steric position for their reaction, as well as catalysing certain acid-base and redox reactions. With the appearance of more complex biomolecules the first specific metalloproteins and among these the metalloenzymes could evolve. The metal ions may stabilize the structure of these macromolecules, but also may participate in their various functions such as electron or oxygen carrier, activator of small inert molecules, or being part of the catalytic active centre. Thus, the properties of metal ions will largely influence the behaviour of the coordinated molecules. In the following few selected examples of the metal ion containing molecules will be listed with their most important bioinorganic chemistry relations.

Probably the most well-known metal ion containing protein in human organism is haemoglobin. Haemoglobin serves as an oxygen carrier. **Fig. 2.** shows the image of a haemoglobin monomer based on its crystal structure. It can be seen that the iron-containing heme group is attached to the protein through a coordinative bond between the metal ion and a histidine side-chain.

Haemoglobin binds the oxygen molecule at the metal ion centre. The cooperative oxygen binding of the four subunits will enable the oxygen binding at high partial oxygen partial pressure, while providing oxygen to myoglobin at low partial oxygen pressure. The basis of the cooperativity lays in the induced movement of the metal ion upon oxygen binding, which induces the conformational change in the neighbouring subunit increasing its oxygen binding ability.

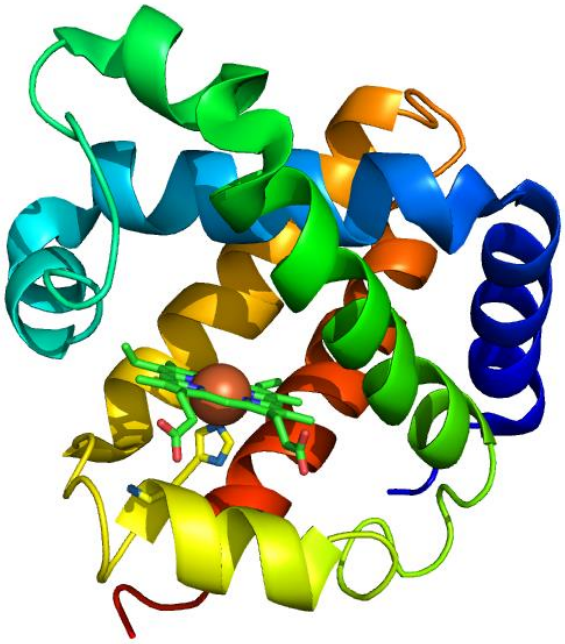


Figure 2. PyMol image of a haemoglobin monomeric based on its crystal structure (coordinates have been downloaded from RCSB Protein Databank, as it has been done for all further PDB files. PDB Id: 1HCO). Heme group is in the centre of the molecule, with an Fe(II) atom as an orange sphere. A histidine side-chain (yellow sticks) is creating the coordinative connection between the protein and the heme group.

Zinc fingers are abundant proteins in living organisms. Zinc(II) ions serve structural role in these proteins. Upon zinc(II) binding the originally unordered protein structure turns into a finger-like ββα motif (**Fig. 3.**). These fingers are able to interact with DNA molecule specifically recognizing selected DNA sequences. As such zinc finger proteins serve mostly as transcription activator factors inducing the synthesis of RNA from its DNA template as the first step of protein synthesis.

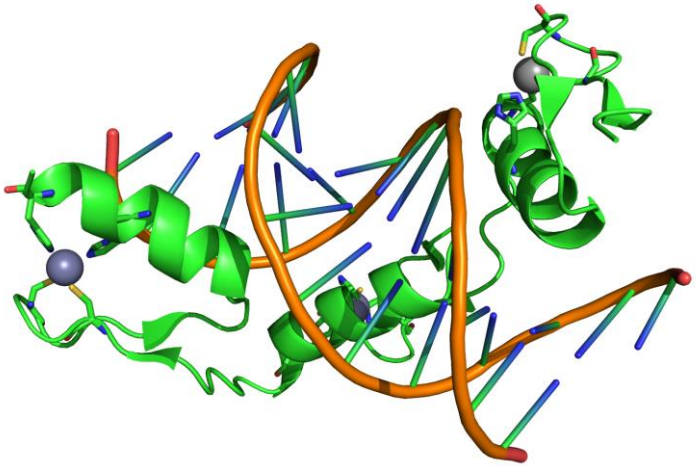


Figure 3. PyMol image of a zinc finger protein based on its crystal structure in complex with DNA (PDB Id: 1MEY). The protein consists of three finger units and is represented by green cartoon. The Zn(II) ions are shown as grey spheres each one coordinated to two histidine and two cysteine side-chains (green sticks). The backbone of the DNA molecule is orange.

The metal ion within a finger unit is coordinated to two imidazole nitrogens in the side-chains of two histidines and two thiolate groups in the side-chains of two cysteines. This type of complex formation saturates the zinc(II) coordination sphere and neutralizes the charge of the metal ion. In such environment zinc(II) does not show catalytic behaviour, such as it does in the active centre of several nucleases. Thus, an unwanted side-reaction, leading to the damage of the DNA during the transcription process is excluded.

Hydrogenases are bacterial enzymes utilizing the redox reaction between the protons and hydrogen as the source of energy or for the storage of energy. Fig. 4. shows a cartoon of a [NiFe] hydrogenase enzyme, which contains an irons sulphur cluster at its active centre containing a nickel(II) ion as well. As the enzymes catalysing redox reactions this protein contains metal ions, which can

change their oxidation states in the active site. The electron transport within the protein structure is ensured by the organized chain of the iron-sulfur clusters. The electron transport is governed by the redox potential of each cluster.

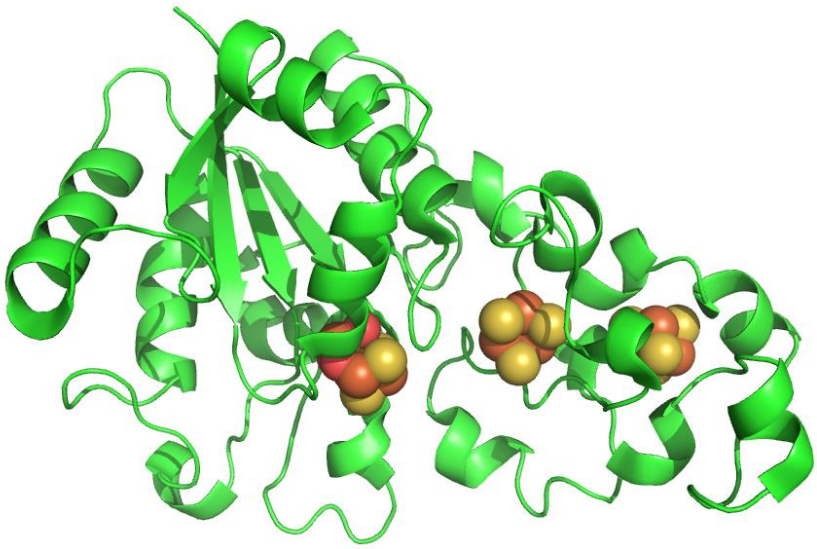


Figure 4. PyMol image of the [NiFe] hydrogenase from *Desulfovibrio desulfuricans* based on its crystal structure (PDB Id: 1E3D). The protein is represented by green cartoon. The metal clusters are shown as spheres.

Metal ions also have a role in unequal electrolyte distribution across the cell membranes with various roles. These phenomena are related to the stability of the cell membrane, changing of the membrane potential leading to the action potential in nerve cells, etc. It is well-known that the concentration of the sodium ion is different at the two sides of the cell membrane. It is much smaller within the cells, than outside. Potassium ions show opposite gradient. These concentration gradients must be maintained for the cell survival. Therefore, the organism uses energy to pump the metal ions against the concentration gradient – a process

called active transport. A Mg^{2+} ion is participating in the catalysis of ATP/ADP conversion (**Fig. 5**).

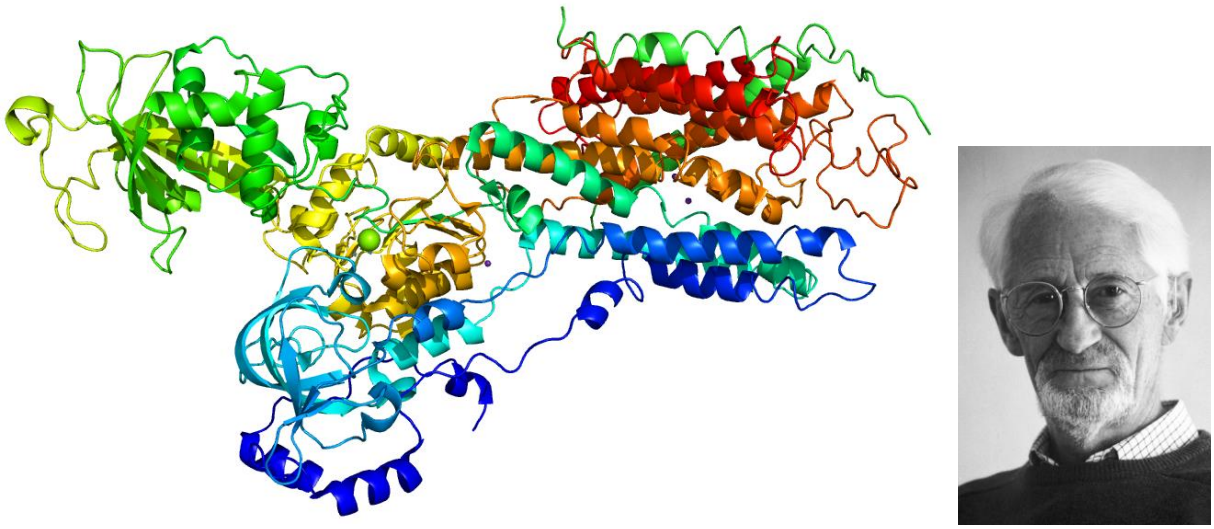


Figure 5. PyMol image of the sodium-potassium pump (left) discovered by Jens Christian Skou (right, photo downloaded from the official website of the Nobel Prize), who received Nobel prize in chemistry for the first discovery of an ion-transporting enzyme, Na^+, K^+-ATPase. The protein is represented by green cartoon based on its crystal structure (PDB Id: 3B8E). The helix bundle at the right side of the structure is the cross-membrane section. Metal ions are shown as spheres.

Numerous further metal ion containing biomolecules can be itemized, but it would be out of the scope of the subject of this e-book. Nevertheless, these examples show the diversity and power of bioinorganic chemistry.

One of the two main approaches of the bioinorganic chemistry is the study of metal complexes and artificial systems modelling the biological structure, function and environment. This approach involves mainly chemical methods of synthesis and study. It has the advantages that the simple/cheap substances

applied in experiments can be synthesized/obtained in a scale that is enough for detailed physicochemical investigations. Simple, well determined systems are studied and straightforward conclusions can be obtained. **Fig. 6.** represents an example from the author's research study on the metal ion peptide system, in which modelling of a nuclease active centre was carried out by a heptapeptide, which coordinates the zinc(II) ion by three histidine side-chain, similarly to many hydrolytic enzymes. This study involved investigation methods commonly used in the modelling in bioinorganic chemistry, such as pH-potentiometric titrations, spectrophotometry and circular dichroism spectroscopy. By these methods we can determine the total concentration of the components, such as e.g. the metal ion, and also the detailed speciation of its different complexed forms under varying conditions. Spectroscopic methods provide opportunity to get insight into the structure of individual complex species, either within the metal ion microenvironment or for the whole molecule. Based on these, one can select appropriate conditions for the further experiments for functional modelling of the studied biological system. In this example, it is shown that the three-histidine coordinated complexes form in the pH range between ~ 5 and ~ 9 (**Fig. 6A**). The result of the DNA cleavage study revealed that at least one of these complexes is catalytically active (**Fig. 6B**).

Modelling is also essential for the development of bioinspired molecules, being able to selectively bind metal ions for metal ion sensing, detecting or accumulating purpose. These agents can be used in various practical applications such as environmental analysis, technology, medicine, etc. The Catalytically active bioinorganic models are promising for the future economic and environmentally friendly technologies.

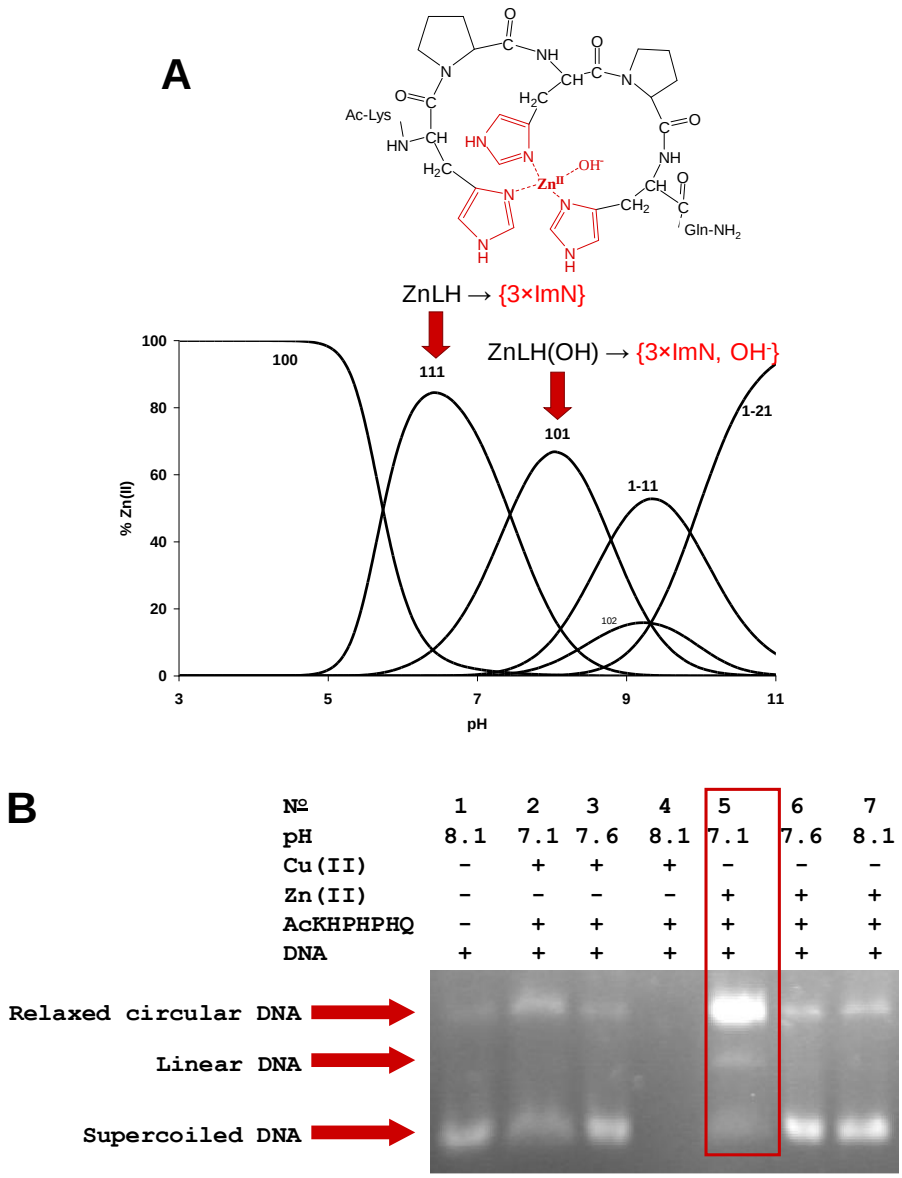


Fig. 6 A) The speciation diagram of the zinc(II)–peptide 1:1.1 system. The scheme of the metal complex, in which the coordination is achieved by three imidazole-nitrogen from the side-chains of histidines is shown above the appropriate species. **B)** The nuclease activity assay of the zinc(II)–peptide complexes is demonstrated by their action on a pUC18 circular DNA. The result of the reaction was visualised by agarose gel electrophoresis – one of the methods discussed later in this e-book. The peptide and zinc(II) together caused an enhanced hydrolytic activity at pH 7.1 demonstrated by the detection of the relaxed and even the linear forms of the DNA in lane 5.

Although we can learn from the modelling experiments, here researchers study isolated systems, and the results are difficult to translate to the real biological environment.

The second main approach is the study of macromolecular systems themselves. These experiments include mainly the metalloproteins and metalloenzymes, but in a broader sense all the metal ion binding biomolecules. These studies will provide direct information on the biologically active substances, but their investigation is difficult with the conventional physicochemical methods. The main reason for this is that the synthesis of the macromolecules poses difficulties. The solid phase peptide synthesis is not suitable for preparation of proteins consisting of hundreds of amino acids. Even the careful synthesis of a 100 amino acid long peptide would yield a mixture of peptides, with similar properties, which would be extremely difficult to purify even using the most advanced separation techniques. Thus, the resulting substance would be too expensive for further experiments.

The Department of Biochemistry at University of Szeged is located in the Institute of Biology. Therefore, chemistry orientated students have only a limited knowledge on this field, which could help them to solve such problems. In this e-book it will be demonstrated that by using the routine methods of the molecular biology the level of the chemical research can be increased. Bacteria can e.g. be used as the most advanced "peptide synthesizers" that can produce proteins with precise sequence. Substitution reactions are often applied in organic chemistry to modify compounds. However, the exchange of a single selected amino acid side-chain within a protein is an almost impossible task by using chemical reactions.

Such modification can be easily executed by the tools of recombinant DNA technology starting out from the gene of the protein as it will be demonstrated later. Similarly, it is impossible to find suitable instrumental separation technique (HPLC or electrophoresis) for the separation of a mixture of DNA molecules with identical length of e.g. 300 bp (where bp stands for the 2'-deoxynucleotide pairs commonly called base pairs). This task can also be easily solved in the knowledge of the rules of the DNA cloning procedure.

In the following chapters, numerous examples of the biological tools will be described and explained, which can be applied in chemistry research. Not only the bioinorganic chemists can utilize these methods, but many others. Counting just few examples:

(i) The drug design e.g. relies essentially on the interactions of the synthetic organic drug molecules with biological macromolecules. Therefore, the drug design is essentially based on the knowledge on the structure and function, and in particular the first step of verification is based on the studies on interactions of these molecules with drugs.

Figure 7. A future photosynthetic car designed by Shanghai Automotive Industry Corporation (source: https://phys.org/news/2010-05-yez-car.html)

(ii) Biosensors are emerging tools of analytical chemistry, as in is demonstrated by the exponentially increasing number of published articles in this field.

(iii) Modern technologies (nitrogen fixation, solar energy utilization (**Fig. 7**), etc.) can arise from the knowledge acquired in the detailed studies and targeted modifications of the natural systems.

Monitoring questions

- How can be a complex research problem in natural sciences solved?

- List examples of overlaps between various research areas in selected research problems.

- What are the roles of metal ions in biological systems? List examples for each type of role.

- What is bioinorganic chemistry?

- What are the methods of studies in bioinorganic chemistry?

- How can be the biological tools applied in chemistry?

3. DNA and the recombinant DNA technology

Students who study this chapter will acquire the following specified learning outcomes:

Knowledge

The students know the basics of recombinant DNA technology.

The students identify the building blocks of the DNA molecule.

The students are aware of the DNA base pairing principles.

The students explain the structure of the DNA double helix.

The students list the main elements of the recombinant DNA technology.

Skills

The students write the sequence of the complementary strand of the DNA based on the knowledge of base pairing.

The students select the complementary (unusual) codes of the nucleobases even if the definition of the nucleobase is not unique.

The students analyse the possibility of the specific and non-specific DNA-protein interactions based on the DNA structure.

Attitude

The students pay attention to the importance of correct writing of the DNA sequence and its complementary sequence.

The students realize the importance of the antiparallel arrangement of the two DNA strands in the double helix.

The students make an effort to realize the importance of the achievements in the history of the DNA research.

Responsibility and autonomy

The students independently search and identify the chemical structures of the DNA components.

The students independently develop their knowledge on the new research on the area of DNA-based drug design.

"Is life just a game where we make up the rules,
While we're searching for something to say,
Or are we just simply spiralling coils,
Of self-replicating DNA?"

Monty Python – The Meaning of Life

The above cited text is the perfect introduction to this chapter defining the recombinant DNA technology (**Fig. 8.**), which is based on the central dogma of the molecular biology.

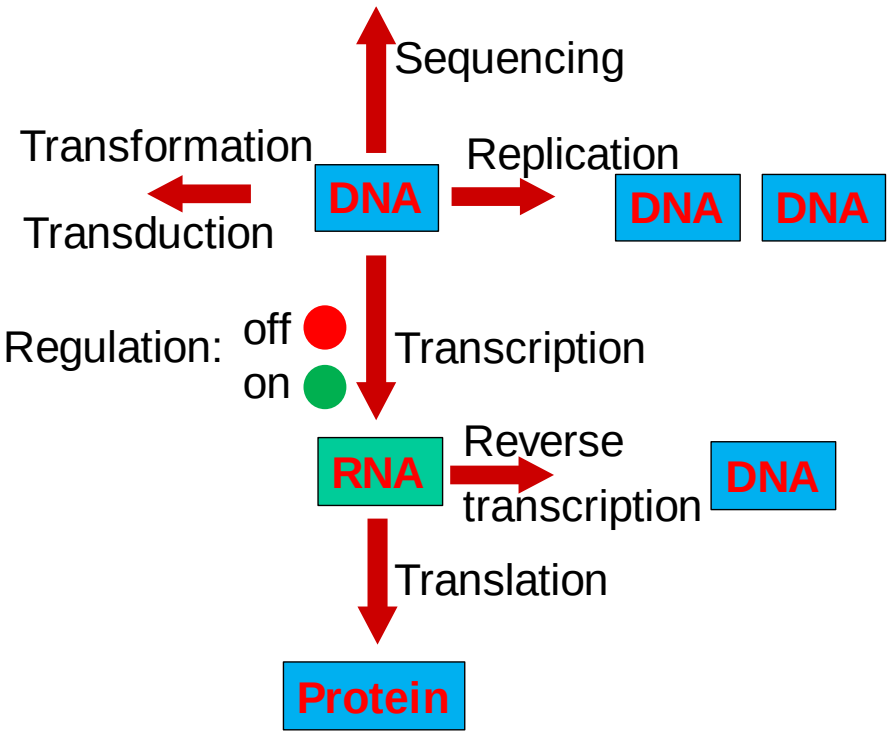


Figure 8. Schematic description of the main processes applied by the recombinant DNA technology.

It means that (i) the information which is stored in DNA is transferred to a newly synthesized DNA molecule in a process called replication; (ii) the information which is stored in DNA is transferred to a newly synthesized RNA molecule in a process called transcription; (iii) the information which is stored in RNA is transferred to newly synthesized protein molecule in a process called translation, being the most abundant processes in the cells. Nevertheless, the transfer of the information in DNA → protein, RNA → RNA, RNA → DNA relations may also occur in nature and have their own importance.

As it can be concluded from **Fig. 8**, the starting point of the protein synthesis – a key molecule in chemistry research – is the DNA. The consequence of this is that the knowledge about DNA should be deep enough to handle such a chemically unfriendly substance. For a long time, researchers could not find suitable tools to better understand its structure and to carry out suitable reactions with it. DNA in the cells is usually a very long and stable polymer consisting of only four (chemically not very) different subunits. This makes it extremely difficult to analyse, synthesize and purify. On the other hand, it is well-known how the recent routine methods of molecular biology revolutionized the knowledge on this research area, as well as the other scientific fields through the interdisciplinary researches. It is very clear that the many discoveries on this field deserved Nobel prizes.

First, the DNA molecule will be introduced in the following. A chemist considers the DNA as a usual chemical identity consisting of atoms, which are bound together with primary and secondary bonds. The building blocks of the DNA polymer are the 2'-deoxyribonucleoside monophosphates (dNMPs), shown in **Fig. 9.** Their common part is the central 2'-deoxyribose, which lacks the

hydroxyl group on the 2'-carbon atom. The hydroxyl group at the 5'-carbon atom is phosphorylated. The glycosidic carbon is substituted by a heteroaromatic compound, a nucleobase, which can either be an adenine, guanine, thymine or cytosine. Accordingly, the individual building blocks are the 2'-deoxyadenosine 5'-monophosphate or simply deoxyadenosine monophosphate (dAMP), deoxyguanosine monophosphate (dGMP), deoxythymidine monophosphate (dTMP) and deoxycytidine monophosphate (dAMP). Adenine and guanine are purine bases, while thymine and cytosine are pyrimidine bases. As an easy research task, find the structural formulae of these molecules.

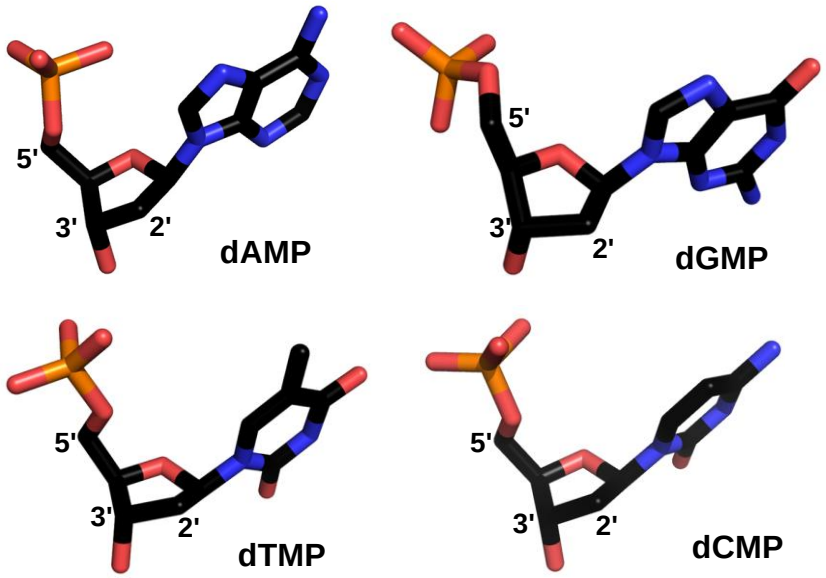


Figure 9. PyMol images of the building blocks of DNA, 2'-deoxyribonucleoside monophosphates from the crystal structure of a DNA molecule (PDB Id: 1ZEW). The abbreviations are explained in the text. Note the numbered carbon atoms of the central 2'-deoxyribose rings, as it will be often referred to these numbers later.

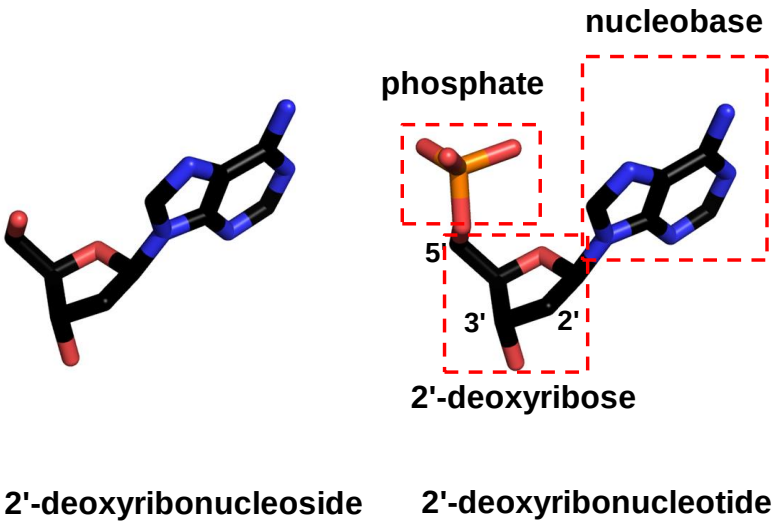


Figure 10. PyMol images of a 2'-deoxyribonucleoside and a 2'-deoxyribonucleotide. The difference between the two molecules is the lack of the phosphate group in the former. The data for the figure have been obtained from the crystal structure of a DNA molecule (PDB Id: 1ZEW).

dNMPs can be more simply named as 2'-deoxynucleotides. Removing the 5'-phosphates the resulting molecules are named as 2'-deoxynucleosides (**Fig. 10.**). The building blocks of DNA are connected together through phosphodiester bonds formed in condensation reaction between the 5'-phosphate and the 3'-hydroxy group of the 2'-deoxyribonucleoside monophosphates (**Fig. 11.**). In fact, soon it will be shown that 2'-deoxyribonucleoside triphosphates (dNTPs) are used as reactive species in DNA synthesis, to provide the necessary energy for the reaction.

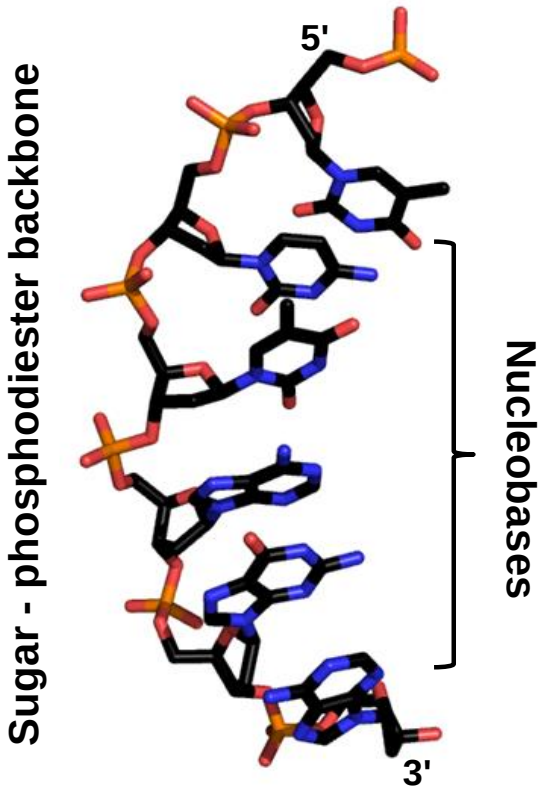


Figure 11. PyMol image of a single strand DNA chain from the crystal structure of a DNA molecule (PDB Id: 1ZEW) showing the phosphodiester bonds forming sugar-phosphate backbone of the DNA molecule. Note the direction of the molecule is determined by the 5'-phosphate and the 3'-hydroxy groups. The text will often refer to the DNA 5' and 3' termini later.

The DNA chain formed in this way, possesses a free 5'-phosphate and a free 3'-hydroxy group (**Fig. 11.**). These groups determine a direction of the DNA chain. The sequence of the DNA is defined as the order of 2'-deoxyribonucleotide units read from the 5'-end of the DNA towards the 3'-end. As an easy exercise try to write the sequence of the DNA chain depicted in **Fig. 11**. For easier reading of the sequence the 2'-deoxyribonucleotides are usually abbreviated with a one-letter code each. These codes refer to the nucleobase (the subunits, which are different among the 2'-deoxyribonucleotide molecules) within each

2'-deoxyribonucleotide. Accordingly, for the 2'-deoxyribonucleotides containing adenine, guanine, thymine or cytosine nucleobases A, G, T or C is written, respectively.

A specific and somewhat surprising property of the single strand DNA molecules (such as it is shown in **Fig. 11.**) that they can bind to a specific second DNA chain to form a double strand DNA. The two DNA chains must be in a specific relationship i.e, they must be complementary to each other to form the double strand. **Fig. 12.** shows how the double strand DNA is formed. It can be seen that the two chains are bound together through multiple hydrogen bonds formed between the nucleobases. It can immediately be recognised that for optimal arrangement of these subunits, the base pairs always consist of one purine and one pyrimidine base. More specifically A is always paired with T and G is always paired with C. It is said that A and T, as well as G and C are complementary to each other. Because of their different size this arrangement provides similar distance between the two chains along the DNA sequence. It is also worth mentioning that the two strands of the double strand DNA have antiparallel arrangement, meaning that the directions of the two strands are opposite to each other. Based on this knowledge try to write the DNA sequence of the complementary strand of the DNA shown in **Fig. 11.**

The two strands form a double helix with a narrow and a broad groove, named minor and major grooves, respectively. This arrangement is also important from the point of view of the interactions of the double strand DNA with macromolecules, such as proteins. Later it will be shown that the proteins recognising specific DNA sequences should interact with nucleobases, and this is only achievable though the major groove in which the nucleobases are exposed to

proteins. Contrary the minor groove is too narrow for a protein to access nucleobases thus, the interactions will mostly be established with the sugar-phosphate backbone, which is uniform along the DNA chain.

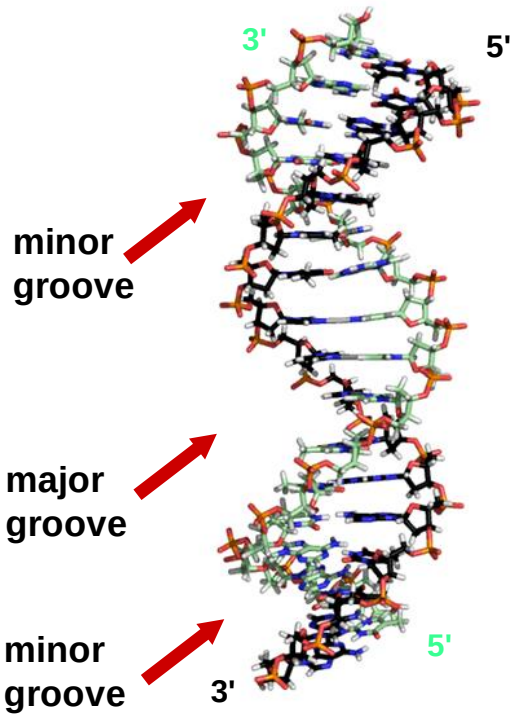


Figure 12. PyMol image of a double strand DNA from its crystal structure (PDB Id: 2M54). Note that the direction of the two single strand DNA molecules is opposite/antiparallel.

It is also important to mention that two hydrogen bonds are formed between A and T and three between G and C (**Fig. 13**). It is not difficult to prevision that the strength of the binding is higher between G and C than between A and T, which will also have consequences concerning the behaviour of DNA in its certain reactions. It is e.g, easier to deform the DNA structure at the A/T-rich motifs than at C/G-rich sequences.

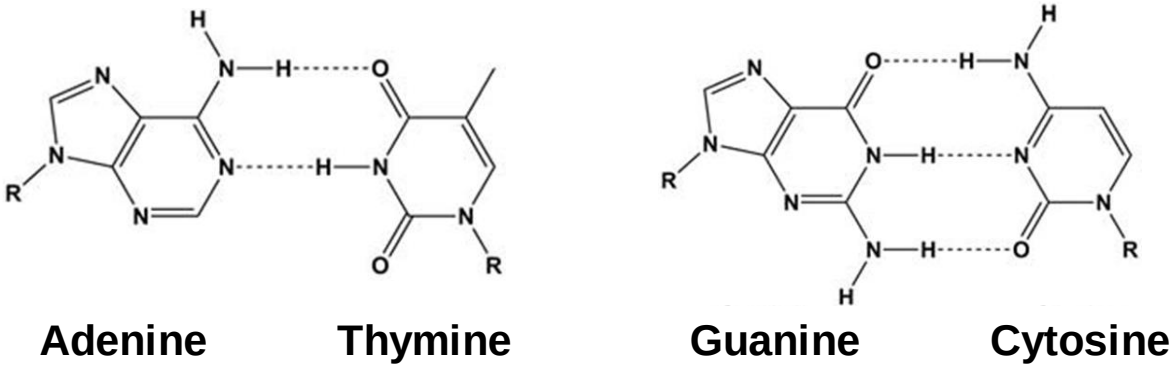


Figure 13. Hydrogen bonds formed between the nucleobases within a double strand DNA.

Table 1. Codes of 2'-deoxynucleotides, including those allowing variation in the sequence or showing ambiguity in the sequence.

Code	Equivalent	Complementary code
A	A	T
C	C	G
G	G	C
T	T	A
M	A or C	K
R	A or G	Y
W	A or T	W
S	C or G	S
Y	C or T	R
K	G or T	M
V	A or C or G	B
H	A or C or T	D
D	A or G or T	H
B	C or G or T	V
X or N	A or C or G or T	N or X

It is worth mentioning that new codes have been introduced for the 2'-deoxnucleotides allowing for the ambiguity in the DNA sequence (**Table 1.**). The significance of these codes will be demonstrated later.

The backbone of the DNA chain is highly negatively charged, as it carries one charge at each phosphodiester bond. For two long DNA chains forming a double strand DNA this results in strong repulsive force. This has to be overcompensated to stabilize the double helix. It is obvious that positively charges ions can shield this high negative charge. Later it will be demonstrated that DNA molecule can be easily precipitated in the presence of salt, which neutralizes the solid substance. However, in addition to the above mentioned hydrogen bonding interactions the hydrophobic interactions between the neighbouring nucleobase residues contribute also to the stabilization of the double strand DNA. The planar nucleobases are arranged to be parallel within the core of the helix and in this way establish stacking interactions.

The knowledge about the DNA listed above was not available until ~ 70 years ago, in spite of the fact that in 1869 Friedrich Miescher isolated a phosphorous containing material from the cell nucleus and published in 1871 (**Fig. 14.**). It was essential to prove that the isolated substance was free of protein impurity, which could turn his conclusions false. He supposed that this "nuclein" is needed for cell division process, and predicted that the knowledge of the interactions between the materials from the nucleus, proteins and their direct metabolic products will gradually enlighten the internal processes of the cells. The experiments were carried out in Felix Hoppe-Seyler’s laboratory in Tübingen. Prior to becoming the chemical laboratory of Tübingen University in 1823, the room was Tübingen castle’s laundry.

The deoxyribonucleic acids then escaped the researchers attention until 1944, when Oswald T. Avery, Colin MacLeod, and Maclyn McCarty suggested that the material responsible for the heredity of various properties in bacteria is DNA instead of proteins, as thought earlier. Not much later the structure of the DNA has been solved.

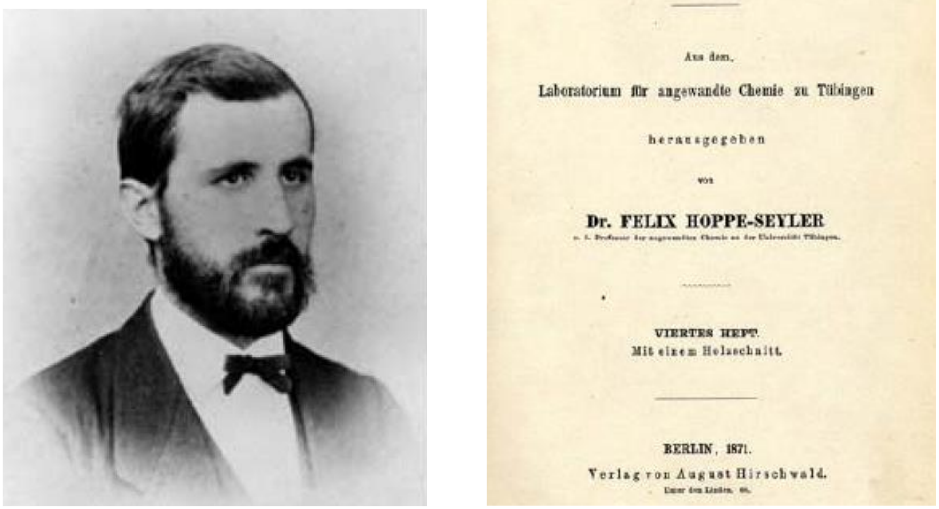
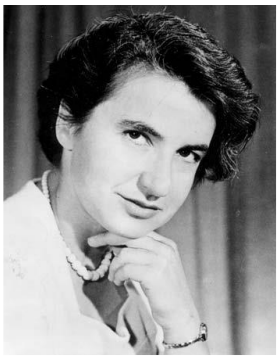


Figure 14. Friedrich Miescher and his paper about the discovery of "nuclein" (DNA) in Hoppe-Seyler's Medicinisch-chemische Untersuchungen (Miescher, F., 1871. Über die chemische Zusammensetzung der Eiterzellen. Med.-Chem. Unters. 4, 441–460).

Rosalind Elsie Franklin (1920 – 1958) was a chemist X-ray crystallographer who contributed to the understanding of the molecular structures of DNA (deoxyribonucleic acid), RNA (ribonucleic acid), viruses, coal, and

graphite. She published her results about the structure of sodium thymonucleate fibres determined by X-ray diffraction in 1953 (**Fig. 15.**).



Acta Cryst. (1953). **6**, 673

The Structure of Sodium Thymonucleate Fibres. I. The Influence of Water Content

BY ROSALIND E. FRANKLIN* AND R. G. GOSLING
Wheatstone Physics Laboratory, King's College, London W.C. 2, England

(*Received* 6 *March* 1953)

Figure 15. Rosalind Elsie Franklin and her paper about the X-ray diffraction experiments in Acta Crystallographica.

Next, James Dewey Watson and Francis Harry Compton Crick published the structure of the DNA double helix in 1953 in Nature (**Fig. 16.**). This achievement deserved a Nobel prize in physiology or medicine in 1962. In the same issue of Nature subsequent articles by Wilkins, Stokes and Wilson, as well as by Franklin and Gosling also discuss the structure of the DNA. Rosalind Elsie Franklin was an X-ray crystallographer, whose X-ray diffraction experiments essentially contributed to the understanding of the molecular structures of DNA Her results leading to the fruitful discussions and finally the discovery of the structure of DNA were largely recognised after her death in 1958. In fact, she might have had the Nobel prize shared with the scientists in Fig. 16, but the Nobel prize is not awarded posthumously. Therefore, she could never receive the deserved highest honour of the scientific community for her work.

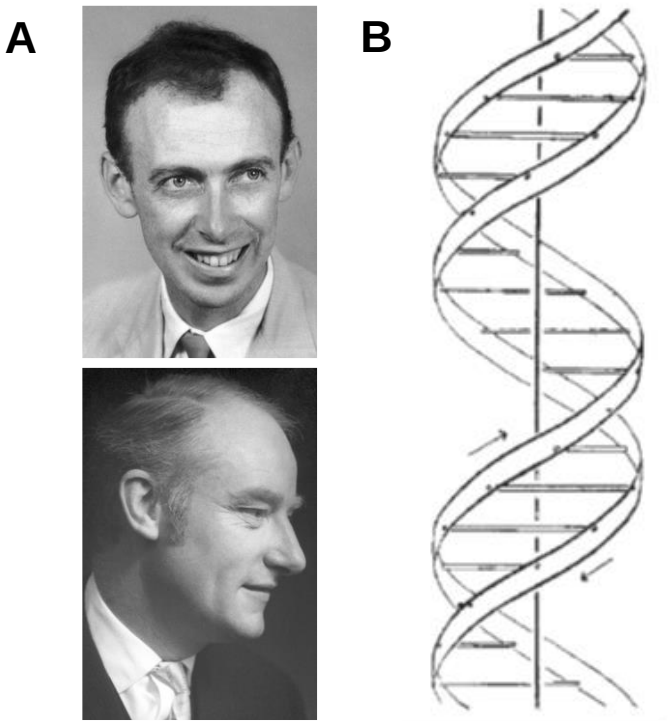


Figure 16. A) From top to bottom: James Dewey Watson and Francis Harry Compton Crick the Nobel prize holders in physiology or medicine 1962 "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material." (Photo from the Nobel Foundation archive.) **B)** The schematic of the DNA structure as published in Nature in April, 1953. The figure was taken from the original article: Nature, 1953, vol. 171, pp.737-738.

The next important step in understanding the properties of the DNA was the cracking of the genetic code in early 1960-s. For this achievement Robert W. Holley, Har Gobind Khorana and Marshall W. Nirenberg received the Nobel Prize in Physiology or Medicine in 1968; while J. Heinrich Matthaei a German biochemist, a post-doctoral visitor in the laboratory of Marshall Warren Nirenberg also significantly contributed to this work. Resolving the genetic code – i.e. the determination of the nucleotide sequences corresponding to the amino acids – was

the most challenging genetic project at that time. The revealed code proved to be essentially the same for nearly all organisms.

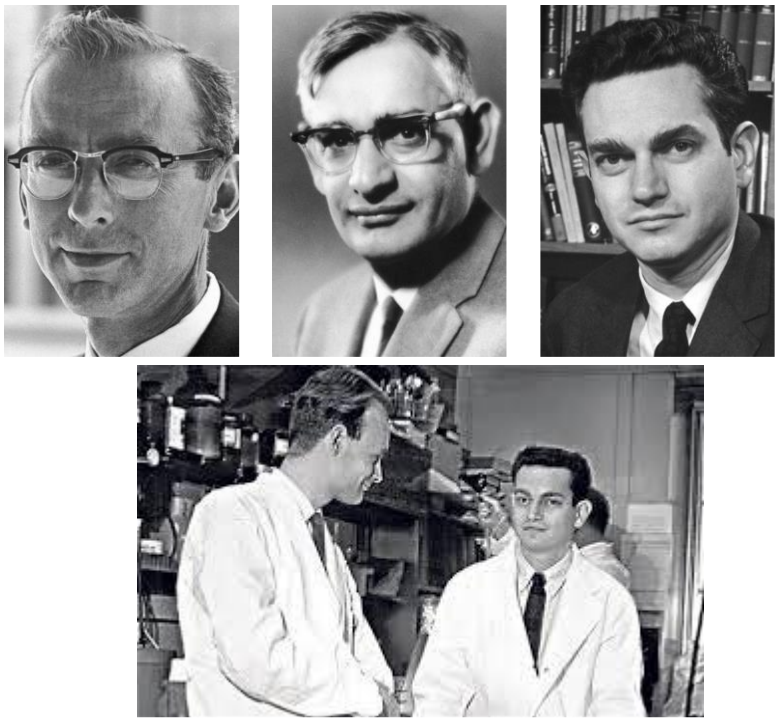


Figure 16. Top left to right: Robert W. Holley, Har Gobind Khorana and Marshall W. Nirenberg the Nobel prize holders in physiology or medicine 1968 "for their interpretation of the genetic code and its function in protein synthesis." (Photo from the Nobel Foundation archive.) Bottom: J. Heinrich Matthaei in Nirenberg's laboratory (source: https://www.telegraph.co.uk/news/science/science-news/8546830/Genes-and-DNA-meet-the-first-man-to-read-the-book-of-life.html).

In the next figure (**Fig. 17.**) the most important milestones of the research results related to DNA are collected in chronological order taken from the excellent review on DNA by Ralf Dahm from the Max Planck Institute for Developmental Biology, Tübingen, Germany.

1865: Gregor Mendel discovers through breeding experiments with peas that traits are inherited based on specific laws (later to be termed "Mendel's laws").

1866: Ernst Haeckel proposes that the nucleus contains the factors responsible for the transmission of hereditary traits.

1869: Friedrich Miescher isolates DNA for the first time.

1871: The first publications describing DNA ("nuclein") by Friedrich Miescher, Felix Hoppe-Seyler, and P. Plósz are printed.

1882: Walther Flemming describes chromosomes and examines their behavior during cell division.

1884–1885: Oscar Hertwig, Albrecht von Kölliker, Eduard Strasburger, and August Weismann independently provide evidence that the cell's nucleus contains the basis for inheritance.

1889: Richard Altmann renames "nuclein" to "nucleic acid."

1900: Carl Correns, Hugo de Vries, and Erich von Tschermak rediscover Mendel's Laws.

1902: Theodor Boveri and Walter Sutton postulate that the heredity units (called "genes" as of 1909) are located on chromosomes.

1902–1909: Archibald Garrod proposes that genetic defects result in the loss of enzymes and hereditary metabolic diseases.

1909: Wilhelm Johannsen uses the word "gene" to describe units of heredity.

1910: Thomas Hunt Morgan uses fruit flies (*Drosophila*) as a model to study heredity and finds the first mutant (*white*) with white eyes.

1913: Alfred Sturtevant and Thomas Hunt Morgan produce the first genetic linkage map (for the fruit fly *Drosophila*).

1928: Frederick Griffith postulates that a "transforming principle" permits properties from one type of bacteria (heat-inactivated virulent *Streptococcus pneumoniae*) to be transferred to another (live nonvirulent *Streptococcus pneumoniae*).

1929: Phoebus Levene identifies the building blocks of DNA, including the four bases adenine (A), cytosine (C), guanine (G), and thymine (T).

1941: George Beadle and Edward Tatum demonstrate that every gene is responsible for the production of an enzyme.

1944: Oswald T. Avery, Colin MacLeod, and Maclyn McCarty demonstrate that Griffith's "transforming principle" is not a protein, but rather DNA, suggesting that DNA may function as the genetic material.

1949: Colette and Roger Vendrely and André Boivin discover that the nuclei of germ cells contain half the amount of DNA that is found in somatic cells. This parallels the reduction in the number of chromosomes during gametogenesis and provides further evidence for the fact that DNA is the genetic material.

1949–1950: Erwin Chargaff finds that the DNA base composition varies between species but determines that within a species the bases in DNA are always present in fixed ratios: the same number of A's as T's and the same number of C's as G's.

1952: Alfred Hershey and Martha Chase use viruses (bacteriophage T2) to confirm DNA as the genetic material by demonstrating that during infection viral DNA enters the bacteria while the viral proteins do not and that this DNA can be found in progeny virus particles.

1953: Rosalind Franklin and Maurice Wilkins use X-ray analyses to demonstrate that DNA has a regularly repeating helical structure.

1953: James Watson and Francis Crick discover the molecular structure of DNA: a double helix in which A always pairs with T, and C always with G.

1956: Arthur Kornberg discovers DNA polymerase, an enzyme that replicates DNA.

1957: Francis Crick proposes the "central dogma" (information in the DNA is translated into proteins through RNA) and speculates that three bases in the DNA always specify one amino acid in a protein.

1958: Matthew Meselson and Franklin Stahl describe how DNA replicates (semiconservative replication).

1961–1966: Robert W. Holley, Har Gobind Khorana, Heinrich Matthaei, Marshall W. Nirenberg, and colleagues crack the genetic code.

1968–1970: Werner Arber, Hamilton Smith, and Daniel Nathans use restriction enzymes to cut DNA in specific places for the first time.

1972: Paul Berg uses restriction enzymes to create the first piece of recombinant DNA.

1977: Frederick Sanger, Allan Maxam, and Walter Gilbert develop methods to sequence DNA.

1982: The first drug (human insulin), based on recombinant DNA, appears on the market.

1983: Kary Mullis invents PCR as a method for amplifying DNA in vitro.

1990: Sequencing of the human genome begins.

1995: First complete sequence of the genome of a free-living organism (the bacterium *Haemophilus influenzae*) is published.

1996: The complete genome sequence of the first eukaryotic organism—the yeast *S. cerevisiae*—is published.

1998: Complete genome sequence of the first multicellular organism—the nematode worm *Caenorhabditis elegans*—is published.

1999: Sequence of the first human chromosome (22) is published.

2000: The complete sequences of the genomes of the fruit fly *Drosophila* and the first plant—*Arabidopsis*—are published.

2001: The complete sequence of the human genome is published.

2002: The complete genome sequence of the first mammalian model organism—the mouse—is published.

Figure 17. The timeline of the milestones in DNA research taken from Developmental Biology, 2005, vol. 278, pp. 274-288.

The figure summarizes a huge success of the researchers in understanding the structure and function of the DNA molecule.

Although the sequence of the human genome (total human DNA) is already known, the research on DNA is far from being finished. Bioinorganic chemists also participate to this research. DNA is e.g, a target of metal ion containing anticancer drugs, because the components of DNA can bind metal ions, which affect its structure and function.

The double strand DNA discussed so far, can also form higher order structures, including its complex with histone proteins. These proteins organize DNA into nucleosomes, which form chromatins, and by further condensation the chromosomes in the human cell nuclei. The double helix of the DNA has a diameter of ~ 2 nm. The distance between to nucleobases is ~ 0.34 nm. Approximately ten nucleotides form on full turn of the DNA strand, i.e, the length of a full turn is ~ 3.4 nm. As an example, the X chromosome in human cells consists of ~ 156 Mbp (millions of base pairs). This size of DNA double helix would be 53 mm long in its linear form. When organized into a chromosome, its dimensions will be ~ 2×8 μm, which can be fit into the cells. The size of human genome is ~ 3200 Mbp, organized into 23 chromosome pairs (the 23rd being the XX for females and XY for males). This long DNA contains short regions, which encode for proteins, as it will be shown later.

It is worth mentioning that the DNA double helix can exist in various conformations. A-DNA, B-DNA, C-DNA, D-DNA, E-DNA, L-DNA, P-DNA, S-DNA, Z-DNA, etc were described so far, but from biological point of view the

first three have most significant appearances. The most common type, described by Watson and Crick is the B-DNA.

In the next chapter the principle of the transfer of the information from a DNA molecule to a newly synthesized DNA molecule will be discussed. The process, called replication is an important element of the recombinant DNA technology.

Monitoring questions

- What are the components of the DNA molecule?

- What is the difference between nucleosides and nucleotides?

- What are the rules of base complementarity and what is the chemistry behind these characteristics?

- Write the complementary sequence of the following DNA strands:

5'-TTAGCCGGTAAGGCCTAT-3'

5'-AGACCMGGTBTAGTGCY-3’

- What is the meaning of the unusual codes in the above DNA sequence?

- How is the double strand DNA built up from the two single strands? List the stabilization and destabilization effects in the double strand DNA.

- Who was awarded Nobel prize for the breaking of the genetic code?

- Characterize the size of the DNA by numbers.

4. DNA replication and the polymerase chain reaction

Students who study this chapter will acquire the following specified learning outcomes:

Knowledge

The students understand the concept of DNA replication.

The students list the enzymes needed for DNA replication within the cells.

The students are aware of the function of the various enzymes in DNA replication.

The students know the theoretical background of PCR

Skills

The students design a PCR experiment listing the materials required to perform such an experiment.

The students analyse the product pattern of a PCR.

The students calculate the amount of the PCR product in each cycle of the reaction.

The students list the possible practical applications of PCR.

Attitude

The students realize the importance of the proper experiment design for PCR and explain this to their colleagues.

The students are motivated to discuss the possibilities of using PCR in their research work.

Responsibility and autonomy

The students design an experimental protocol for PCR on their own.

The students realize the necessity of the responsible evaluation of the PCR results in practice, such as inherited diseases, parenthood affiliation, criminalistics, etc.

The students independently study about the further opportunities of PCR

The complementarity of the nucleobases within the DNA sequence guarantees that the nucleotide sequence of the parent DNA double helix i.e, the genetic information, will be precisely copied into the newly synthesized molecule. The nucleotide sequence in one strand of the double strand DNA determines the nucleotide sequence of the complementary strand. Therefore, both strands of the DNA can serve as the template for the DNA synthesis, resulting in the same double strand DNA product. This process is visualized schematically in **Fig. 18**. As an exercise, identify the nucleobases in the figure.

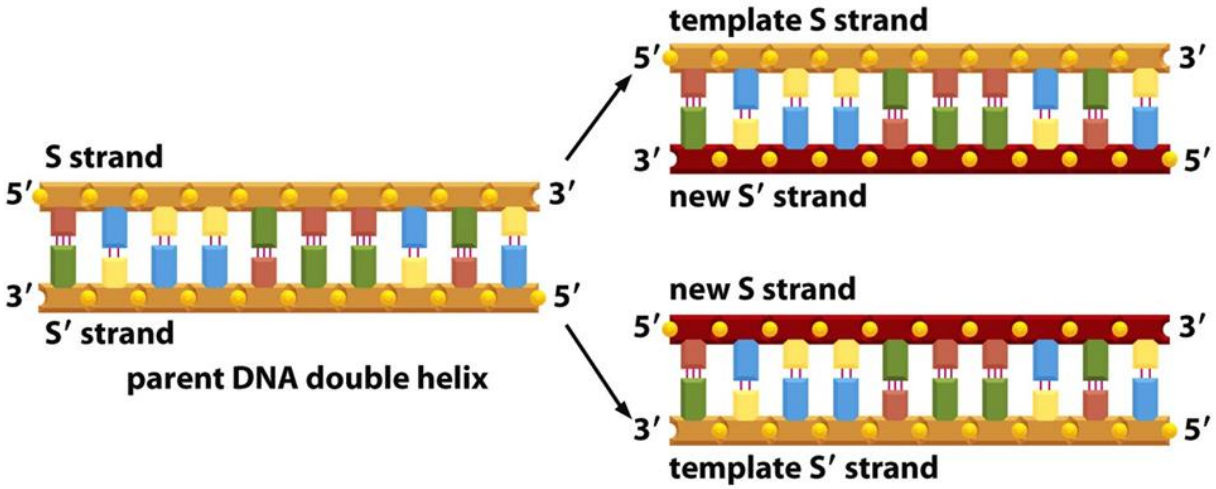


Figure 18. The schematic representation of the duplication of the DNA molecule. The figure is taken from The molecular biology of the cell, Garland Publishing Inc, New York, London, 1989.

This seemingly simple process utilizes a large number of proteins/enzymes in living cells. In the following the focus will be set strictly on the replication process. First the replication in the cell will be summarized briefly. The long double strand DNA molecule has to be separated into single strand DNA

molecules behaving as templates in the area of the replication process. For this the cells apply the helicase enzymes. These enzymes use the energy of adenosine triphosphate (ATP) to separate the two DNA strands. The crystal structure of helicase isolated from T7 phage shows that the molecule consists of six subunits, but it is not sixfold symmetric (**Fig. 19.**). The conformation of the hexamer depends on the form of the substrate bound. In opposite positions pairwise it is either ATP or ADP bound or the substrate binding site is empty. The three sites interconvert in the coordinate fashion through conformational changes, causing a rotation-like movement and oscillation of the enzyme resulting in unwinding the double strand of the DNA. Close to the helicase a replication fork is formed.

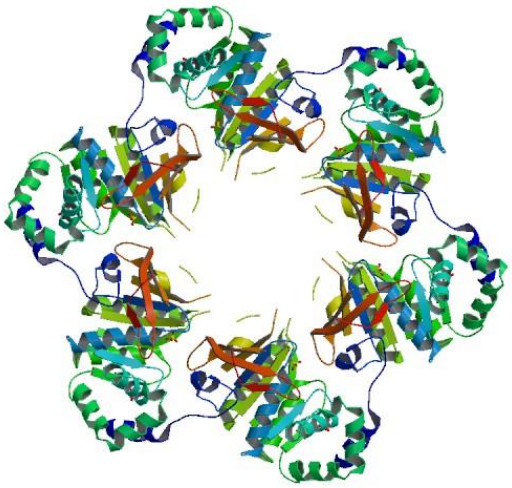


Figure 19. The crystal structure of the helicase enzyme isolated from T7 phage (PDB Id: 1CR0).

The new DNA strand is built up by another enzyme, called DNA polymerase. However, this enzyme can only build up the new strand starting from an existing initial sequence. This sequence, further denoted as primer, is

synthesized by the DNA primase enzyme. From this sequence the DNA polymerase starts to synthesize the new strand from the appropriate energy-rich building blocks, the 2'-deoxyribonucleoside triphosphates (dNTPs). The 5'-triphosphate group will be the reactive side of the new building blocks, which will react with the free 3'-hydroxy group of the primer or the new growing strand, if the correct base pairing is established. This means that the new strand is always growing in the 5' → 3' direction as shown in **Fig. 20**.

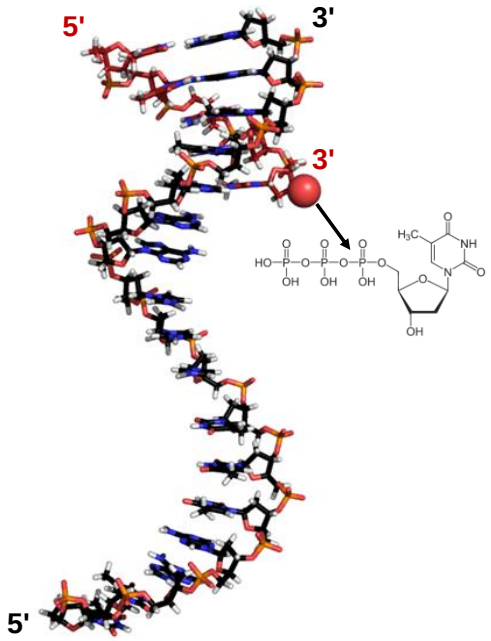


Figure 20. The crystal structure of the template strand DNA with a primer sequence (with red backbone) hybridized to the 3' terminus of the template strand. In this way the 3' terminus (symbolized by red ball) of the primer is directed toward the template, thus it can react with the incoming dNTP, by the help of the DNA polymerase enzyme. The primer is thus, prolonged by stepwise attaching the incoming dNTPs to the 3'-OH. A pyrophosphate is released during each phosphodiester bond formation. (PDB Id: 2M54).

An important consequence of this rule is that following the helicase, one of the two DNA polymerases can continuously build the new strand at the template strand directed with its 5' end to the polymerase – called leading strand template. The other template strand, however, has opposite direction – called lagging strand template –, which requires the second DNA polymerase to build the new strand in backward direction compared to the path of the helicase. As the consequence of this, the synthesis of the second strand very complicated (see **Fig. 21.**). Further proteins and enzymes participate in it, and the new strand is synthesized piece by piece.

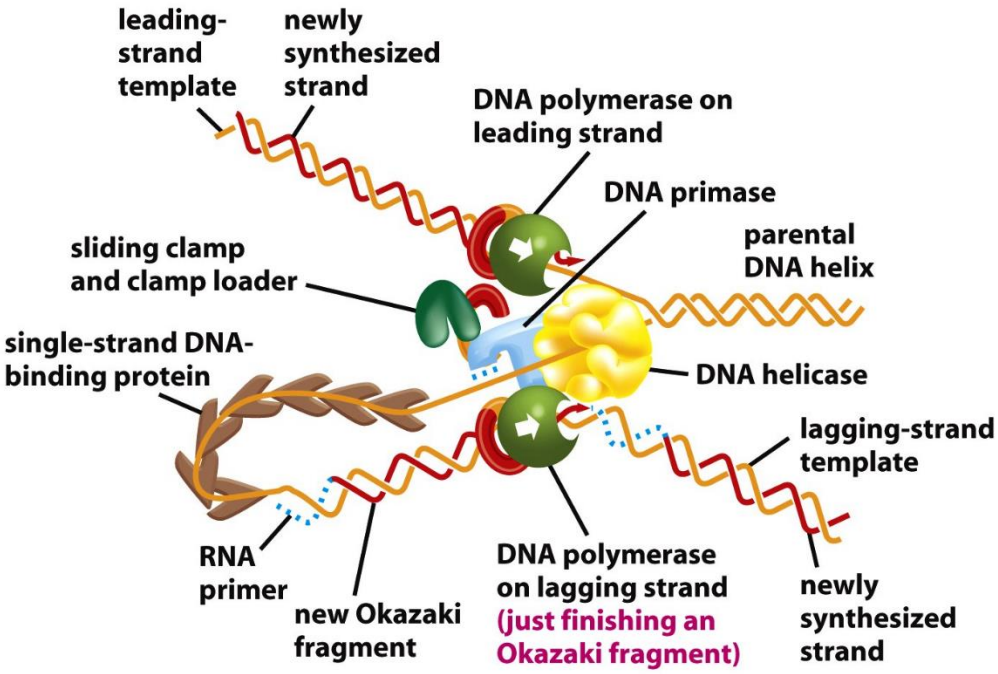


Figure 21. The schematic representation of the DNA replication fork with the protein machinery attached to the template strands. The growing new strands are in red, while the template strands are orange. The figure is taken from The molecular biology of the cell, Garland Publishing Inc, New York, London, 1989.

The new DNA fragments, called Okazaki fragments which are finally linked together by the help of a ligase enzyme. Nevertheless, the process is well optimized and it is so efficient, that the two strands can grow in parallel.

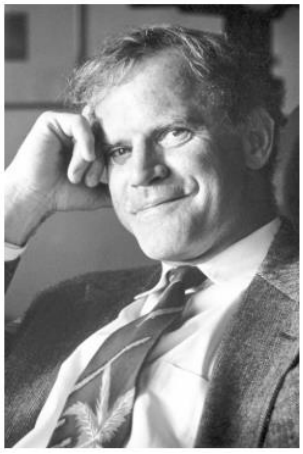
The DNA replication is extremely precise and quick process in the cells. The helicase is rotating with approximately the speed of a jet engine. The polymerase enzyme catalyzes the formation of ~ 5000-10000 phosphodiester bonds in one minute, and allows on average for only one mistaken base pairing in every 10000^{th} case. Some DNA polymerases possess so-called proofreading activity, by means of which they can correct even this error in the new DNA sequence. By means of this safety function they allow erroneous base pairing only in every 10^8 nucleotide unit.

The results of the research on the field of the DNA replication conducted in the laboratories of Max Delbrück, Alfred D. Hershey and Salvador E. Luria also deserved Nobel prize (**Fig. 22.**).

Figure 22. The Nobel Prize in Physiology or Medicine 1969 was awarded jointly to Max Delbrück, Alfred D. Hershey and Salvador E. Luria (from left to right) "for their discoveries concerning the replication mechanism and the genetic structure of viruses." (Photo from the Nobel Foundation archive.)

The DNA replication within the cells is often used to multiply DNA in recombinant DNA technology, but in this way only the DNA recognized by the cell as its own component will be multiplied. Elaborating the method of efficient DNA replication in a test tube would allow researchers to multiply any desired DNA sequence. Certainly, the final goal of these experiments is to produce recombinant proteins from recombinant DNA molecules. By this technology, the modification of proteins, including mutation, truncation, fusion to other proteins would become feasible. The first step of this is to synthesize and modify DNA sequences coding for these proteins. Recombinant DNA molecules are DNA segments from different biological sources combined together to obtain new genetic material. Such recombinant DNA could be a bacterial DNA including a target gene for protein expression in bacterial cell.

To replicate the DNA in a test tube, it is necessary to avoid the use of various enzymes, since this would make the reaction very expensive. It is possible to separate the two strands of any template DNA by increasing the temperature of a reaction mixture to ~ 100°C. This, however, would be deleterious for the DNA polymerase, the presence of which is the minimal requirement for DNA synthesis. The solution to this problem was the discovery of the heat-resistant DNA polymerase (Taq polymerase) isolated from the *Thermus Aquaticus* extremophyl bacterial species living in hot springs and geysers. This enzyme can survive the increased temperature for separating the double strand of the DNA. It is worth mentioning that Science nominated the Taq polymerase enzyme as the molecule of the year in 1989. The other brick in the wall was the brilliant idea of Karry B. Mullis (**Fig. 23.**), working as a chemist at Cetus Corporation in USA.



Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction

K. MULLIS, F. FALOONA, S. SCHARF, R. SAIKI, G. HORN, AND H. ERLICH
Cetus Corporation, Department of Human Genetics, Emeryville, California 94608

Figure 23. Kary Banks Mullis (1944-2019) obtained Nobel Prize in Chemistry in 1993 "for contributions to the developments of methods within DNA-based chemistry - for his invention of the polymerase chain reaction (PCR) method", and his scientific paper describing the method. (Photo from the Nobel Foundation archive.)

He introduced a pair of primers to determine the termini of the new DNA molecule to be amplified in subsequent cycles, periodically changing the temperature. Repeating the cycles of separating the double strand DNA into single strand templates, the hybridization of the primers a lover temperature, and the new DNA strand prolongation by the help of the DNA polymerase enzyme, the exponential amplification of the target DNA has been achieved. The method has been published in 1986 in a paper in Cold Spring Harb Symp Quant Biol, vol. 51, pp. 263-273. These discoveries lead to the polymerase chain reaction (PCR) and the Nobel Prize for its invention in chemistry in 1993. The PCR process revolutionized the fields of biochemistry and molecular biology, and its invention had also strong economical consequences: the Cetus company, as the owner of

the right of the technique from 1989, has sold the patent of PCR and Taq polymerase in 1992 to Hoffmann-LaRoche company for 300,000,000 USD.

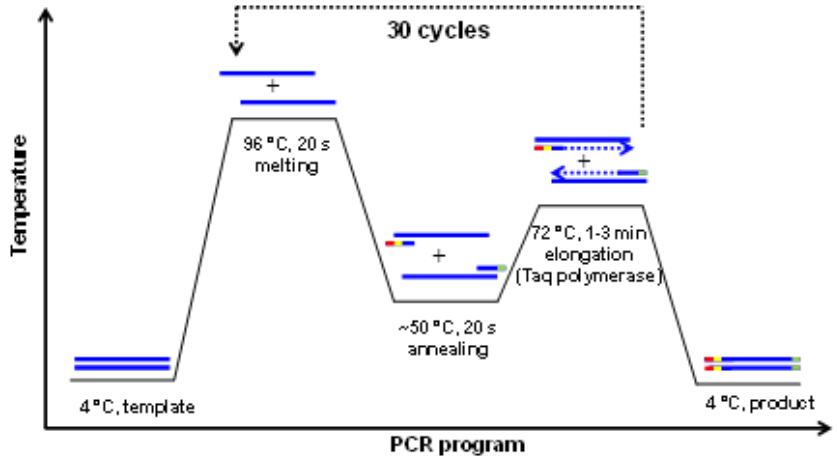
If the genome carrying the gene of the target protein is available, selection of the gene, and its mutations can be carried out by polymerase chain reaction. The PCR reaction mixture contains the following components:

- template DNA molecules;
- the heat stable polymerase enzyme in a Mg^{2+}-containing buffer, e.g. Taq polymerase;
- DNA primers: short synthetic 2'-deoxyoligonucleotides that can hybridize to the single strand DNA template and allow the polymerase to start the synthesis of the complementary strand;
- 2'-deoxynucleoside triphosphates (dNTPs: dCTP, dATP, dGTP and dTTP) as building blocks.

The reaction is driven by heat control (**Fig. 24A.**). In the first step of the cycle, the denaturation, the reaction mixture is heated to > 90 °C. The template DNA dissociates to single strands at around the so-called melting temperature. Then the temperature is lowered to a point close to the melting point of the primers (usually ~50-65 °C) so that they can hybridize to the single stranded template (annealing). In the third step of the cycle the temperature is adjusted to the optimum for the polymerase enzyme function, which is ~72 °C in case of the Taq polymerase, and the elongation of the primers towards the 3' end of the new strand using the dNTPs as building blocks results in the new double strand DNA molecule. These three steps are then repeated in the PCR instrument (**Fig. 24B.**). This instrument is adjusting the preprogrammed temperature very precisely and

the temperature change is very quick. It is essential to use special thin-walled test tubes for efficient heat transfer.

A



B

Figure 24. A) The temperature program of a PCR used to introduce mutation to the end of the sequence and scheme of the reaction. Each blue lines indicate a strand of the template DNA. The red and green endings are restriction enzyme cleavage sites, carried by the primer molecules. The mutation introduced by primers is in yellow. These will be explained later in more detail in the section dealing with the design of the primers. (Taken from the PhD dissertation of Eszter Németh, written in the laboratory of the author of this e-book.) **B)** The PCR instrument used in the laboratory of the author.

The products of each cycle formed in the elongation step serve as templates in the next cycles. The amount of the PCR products is mathematically doubled in each cycle. Thus, the amount of DNA is exponentially increasing: the number of products from one template molecule will be ~2^n where n is the number of cycles.

As already mentioned the primers determine the termini of the newly amplified DNA molecule. Thus the termini can be selected by the appropriate selection of the primer sequences. By means of this, any target gene can be selected from a genome for amplification (**Fig. 25.**).

At the beginning of the reaction from the original template strands longer PCR products are formed than the desired ones, as there is no barrier for the polymerase to stop the process. The prolongation will be terminated by the increase of the temperature at the beginning of the next cycle. From the next cycles the PCR products formed in the previous cycles will also serve as templates. The PCR from these templates will yield the product with the precisely the selected length – now determined by the end of the new template strands (see **Fig. 25.**). The "long products" can for only from the original template, i.e. their amount is increasing linearly. Therefore, after 20-25 cycles these can be neglected in comparison to the exponentially increasing amount of the selected DNA section. In the 25^{th} cycle starting out from a single DNA molecule we can obtain 50 single strand DNA molecules of undetermined length, and ~ 33.55 millions of the targeted double strand DNA with the desired length. Thus, the selected section of the DNA is efficiently multiplied in PCR.

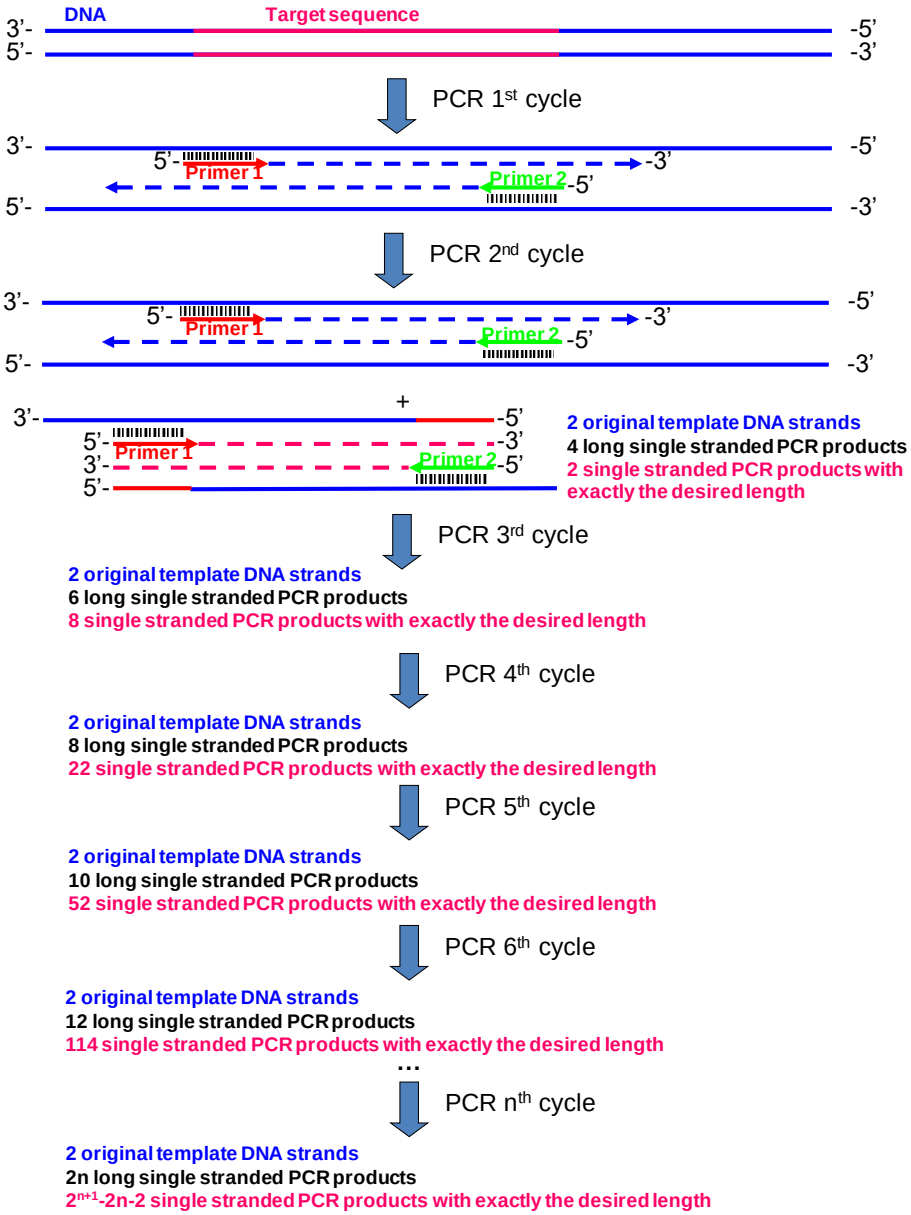


Figure 25. The schematic representation of the PCR with a single DNA template molecule. The blue strand form the original templates, form which the pink target sequence is selected by the primer pairs. The primer shown in red is usually called forward primer, while the red one is the reverse primer. (Other designations such as 5' and 3' of N-terminal and C-terminal primers, respectively may be found in the literature.)

It is very important to adjust the proper temperature in each step of the cycle. The first is usually a high temperature above 90 °C, as this temperature is high enough to separate the strands of almost any DNA. To be sure that this occurs at the beginning of the reaction an extra denaturation step is usually inserted before the first cycle at >90 °C for few minutes. In the next cycles usually short from few hundred up to few thousands of base pairs) selected target DNA sequences appear as double strands, which can be separated within a short time. Thus, the time of the denaturation in the cycles can be between 15-30 seconds. In the second step of the cycle the temperature is decreased to allow the primers to hybridize to their complementary sequences at the template strands. The temperature here has to be selected very carefully. Too high temperature compared to the melting point of the primers can prevent the hybridization of the primers, and thus the amplification of the DNA. In contrast, too low temperature, may cause the hybridization of the primers to sequences, which are similar to their complementary ones. This may result in parallel amplification of more fragments, some of which will be different from the desired sequence. According to the experience of the author, the first trial should be carried out at the temperature, which is higher than the calculated melting point of the primer by ~ 5 °C (the reason for this is most probably the presence of various cations in the reaction mixture which promote the hybridization shielding the negative charges of the sugar-phosphate backbone). The methods of calculation of melting point will be detailed in the next chapter. The time for the annealing step is usually 15-30 seconds. As it was already mentioned, the temperature of the prolongation step shall be adjusted to the optimum of the DNA polymerase function. Different polymerases require different temperature, which is provided by the supplier. It is

usually between 68 and 72 °C. The time for this step depends on the length of the selected DNA sequence for the amplification. As a general rule, the DNA polymerase under the conditions of the PCR can build up a DNA sequence of approximate length of 1000 nucleotides within one minute. Accordingly, the necessary time can be calculated.

Fig. 26. shows an example of a successful amplification of some selected DNA molecules in a PCR performed in the instrument shown in **Fig 24B**.

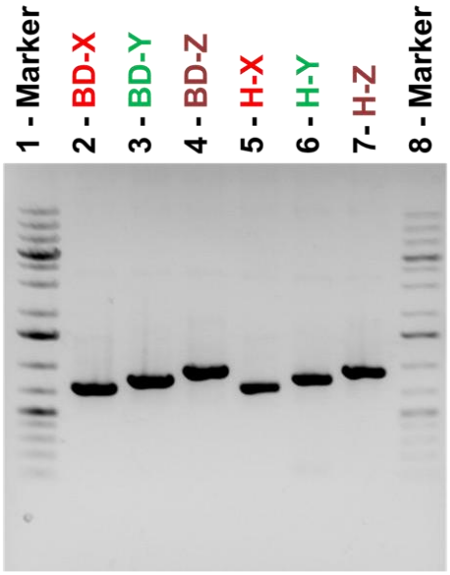


Figure 26. Visualization of the results of PCR by agarose gel electrophoresis. The details of the technique will be discussed later. Here lanes 2-7 show the bands of the desired PCR products, in comparison with a mixture of DNA molecules of known size (lanes 1 and 8). The most important conclusion from this figure is that unique bands were detected in each lanes of the PCR products, showing that there is a unique major product. The band of the linearly amplified "longer fragments" and the bands of the template DNA are not visible, as expected from the description of the PCR in the text.

After around 30 cycles the efficiency of the enzyme starts to decrease. One of the many reasons is that the high temperature and the temperature changes will slowly inactivate the enzyme under the experimental conditions. Therefore, it is not advisable to adjust the number of the cycles to higher than 30. Also the amount of the primers and the dNTPs is decreasing during the reaction as these are built in the newly synthesized DNA sequences, the number of which is exponentially growing. The more original template we apply at the beginning of the reaction the more quickly the primers and the dNTPs are consumed. Thus, high amounts of the templates will result in the quick decrease of the materials needed for the DNA synthesis. This might also be the situation when amplifying long DNA fragments. The PCR might be suitable for amplification of up to 20 kbp long DNA fragments, but for such reactions specially optimized long PCR polymerase enzymes are needed usually with a proof reading activity.

This means that the DNA synthesis will be stopped at low cycle number. The lower is the cycle number the lower is the ratio of the number of the desired products and those with undefined length and the template. The dNTPs are usually added in equivalent amounts each. Statistically this is a good choice, since in a long DNA template the four nucleotides occur with almost the same probability. However, if the selected DNA sequence for the amplification shows an uneven distribution of the various nucleotides, this has to be taken into account when constructing the reaction mixture.

In some specific cases the amount of the DNA template may be very small. This is often the situation when using human DNA samples. To achieve reasonable amplification of the selected gene from such a template the PCR is carried out as usually, but an aliquot of the product mixture resulting from the first

reaction will serve as the template in a second PCR. In this way we multiply the template in the first PCR, which used as the template will result in efficient second PCR. This procedure is sometimes called nested PCR.

By means of this strategy very sensitive method is obtained able even to identify the origin of a single DNA molecule. The presence or absence of the amplified DNA in a properly executed PCR may provide information about the presence, or about the altered sequence of the hypothesized template in the reaction mixture. Therefore, such experiments can also be used in various fields other than the recombinant DNA technology, such as e.g. the examinations of the inherited diseases, the parenthood affiliation tests, the criminalistics to identify the origin of the DNA found at crime scene, etc.

As an example the procedure of the finding of the inherited genes is described in the following. There are homologous short tandem repeat (STR) DNA sections in human genome. Commonly a repeat unit consists of 2–7 base-pairs. These sequences are polymorphic, and they are situated in non-coding regions of DNA. According to the present knowledge, they do not have a specific function in the organism. Nevertheless, they are of great importance for us, since their pattern is characteristic for every person. Thus, their investigation e.g. by PCR can be applied for identification of a person. For this reason, the method is called "DNA-fingerprint"; and today the technique is often applied in various investigations. Notice that in this PCR several DNA fragments are amplified at the same time in the same reaction mixture. Is such experiments the composition of the reaction mixture has to be optimized with respect to the individual primer pairs, the building blocks and also the components of the buffer solution.

The extraordinary sensitivity of the PCR made it an important tool for such investigations. The STR loci are multiplied using specific primers designed to hybridize to their termini. Then the PCR products of different sizes can be separated from each other by a capillary electrophoretic method with very good resolution. If the primers are labelled with e.g, different fluorescent dyes several sections can be studied simultaneously. The comparison of the electrophoretic pattern of the samples obtain from different sources the identification can easily be ascertained.

Monitoring questions

- Describe, the basic concept of DNA replication?

- How the DNA molecule is replicated in the cells?

- Which enzymes participate in the DNA replication within the cell?

- What is the difference between the replication of the leading and the lagging template DNA strands?

- Is there such a phenomenon in PCR as described in the previous question?

- How the amplification of a DNA molecule is carried out in PCR?

- What are the constituents of a PCR reaction mixture?

- Which DNA polymerase is suitable for PCR and why?

- How many cycles is advisable to perform in a PCR and why?

- Who was awarded Nobel prize for the invention of PCR?

- What can be PCR used for in the practice?

5. Primer design for the polymerase chain reaction

Students who study this chapter will acquire the following specified learning outcomes:

Knowledge

The students understand the concepts of primer design.

The students know the precise role of the primers in PCR.

The students know the basics of the primer optimization.

The students are aware of the methods of melting point calculation.

Skills

The students critically evaluate the primer design based on the various aspects.

The students optimize the primer length based on the hybridization properties.

The students collect the opportunities of the gene modification using various primers and various strategies in PCR

The students are capable of primer design for various purposes.

Attitude

The students pay attention to the importance of correct design of the oligonucleotide primers.

The students make effort to optimize the primers from various aspects, aiming at an economic experiment.

The students realize the importance of the correct writing of the DNA sequence in their oligonucleotide order forms.

The students are critical when selecting the strategy of the gene modification by PCR.

Responsibility and autonomy

The students are aware of the high costs of the biomolecules, and chose the most economic experiment strategy.

The students optimize the PCR primers independently.

The students discuss with their colleagues about the opportunities of gene modification using PCR.

The students independently improve their knowledge on the field of primer design.

Primer design is a key step of the PCR. The primers are synthetic 2'-deoxyoligonucleotides, which hybridize to the 3'-end of the selected gene within the template DNA strand, so that they can be prolonged starting from their 3'-hydroxyl groups. In the simplest case of PCR, i.e. selecting a certain DNA segment of the existing DNA template the termini of the target DNA are well defined. Thus, the primer sequences have to be chosen to be complementary to the 3' end of this sequence, in other words, they can be read from the template strands (**Fig. 27.**). There is no opportunity here to change the sequence of the primer compared to the template, unless the introduction of a mutation is the goal of the experiment.

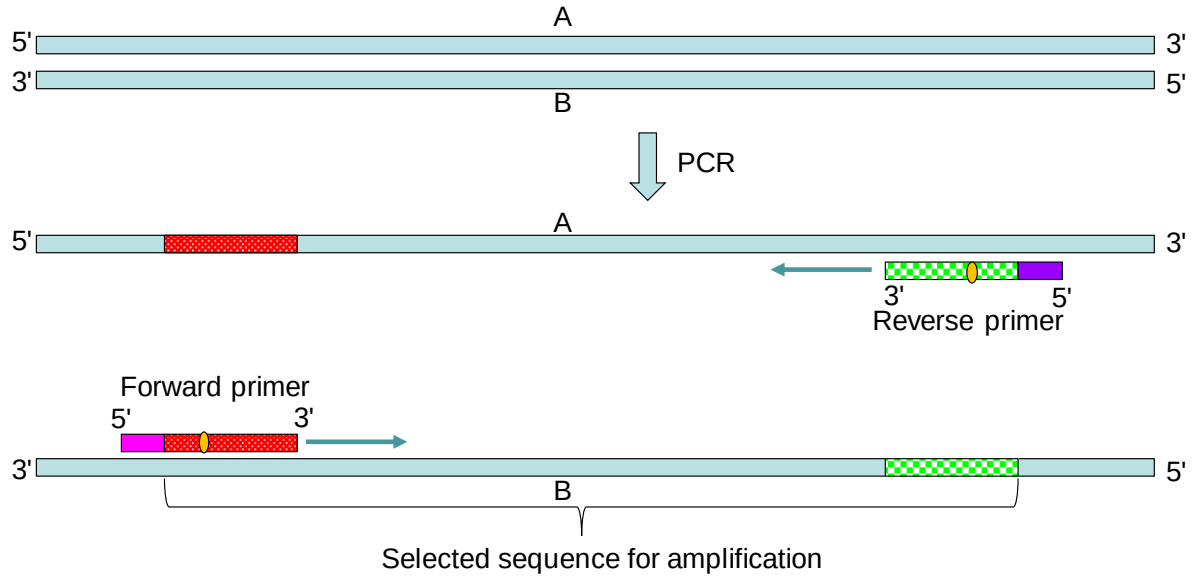


Figure 27. The schematic representation of the hybridization of the primers to the template strands in PCR. The sequence of the forward primer can be read from the A strand, while that of the reverse primer from the B strand as shown in the figure.

In the following, an example of such primer design will be considered. The part of the genomic DNA is shown below, but only the strand from which the coding sequence can be directly read – suppose it is strand A in **Fig. 27**. As mentioned above the sequence of this strand determines the sequence of the complementary strand. The target sequence to be amplified in PCR is bold italic.

5′-GGTTCTCCTCCCTCCTCCTCGCATTCTCCTCCTCCTCTGCTCCTCCCGATCCCTCCTCC
GCCGCCTGGTCCCTCCTCCTCCCGCCCTGCCTCCCCGCGCCTCGGCCCGCGCGAGCTAGACG
TCCGGGCAGC***CCCCCGCCGCAGCCCCG******GCCGCAGCAGCCTCCGCCCCCCGCACGGTGTGAGCGC***
CCGACGCGGCCGAGGCGGCCGGAGTCCCGAGCTAGCCCCGGCGGCCGCCGCCGCCCAGACCG
GACGACAGGCCACCTCGTCGGCGTCCGCCCGAGTCCCCGCCTCGCCGCCAACGCCACAACCA
CCGCGCACGGCCCCCTGACTCCGTCCAGTATTGATCGGGAGAGCCGGAGCGAGCTCTTCGGG
GAGCAGCGATGCGACCCTCCGGGACGGCCGGGGCAGCGCTCCTGGCGCTGCTGGCTGCGCTC
TGCCCGGCGA**GTCGGGCT**CTGCAGCAAAACAAA**GGTAAGGGCGTGTCTCGCCGGCTCCCGCG
CCGCCCCCGGATCGCGCCCCGGACCCCGCAGCCCGCCCAACCGCGCACCGGCGCACCGGCTC
GGCGCCCGCGCCCCCGCCCGTCCTTTCCTGTTTCCTTGAGATCAG-3′

To carry out a successful PCR, first the design of the primers has to be accomplished. The forward primer can directly be read from the 5' end of the target sequence. The selected forward primer is highlighted by red background. It consists of 15 nucleotides. The reverse primer can be selected from the strand B. The sequence of this strand can be obtained by the inversion of the strand A. The inverse sequence of a DNA corresponds to its complementary and reverse sequence. For a long DNA it is time consuming to convert the sequence by hand, but many programs are available also on-line (an example can be The Bio-Web website: http://www.cellbiol.com/cgi-bin/complement/rev_comp.cgi). In this way the DNA sequence of chain B in **Fig. 27.** can be written starting from its 5' end.

5′-CTGATCTCAAGGAAACAGGAAAGGACGGGCGGGGGCGCGGGCGCCGAGCCGGTGCGCCG
GTGCGCGGTTGGGCGGGCTGCGGGGTCCGGGGCGCGATCCGGGGGCGGCGCGGGAGCCGGCG
AGACACGCCCTTACC***TTTCTTTTCCTCCAGAGCCCGACTCGCCGGGCAGAGCGCAGCCAGCA***
GCGCCAGGAGCGCTGCCCCGGCCGTCCCGGAGGGTCGCATCGCTGCTCCCCGAAGAGCTCGC
TCCGGCTCTCCCGATCAATACTGGACGGAGTCAGGGGGCCGTGCGCGGTGGTTGTGGCGTTG
GCGGCGAGGCGGGGACTCGGGCGGACGCCGACGAGGTGGCCTGTCGTCCGGTCTGGGCGGCG
GCGGCCGCCGGGGCTAGCTCGGGACTCCGGCCGCCTCGGCCGCGTCGGGCGCTCACACCGTG
CGGGGGGCGGAGGCTGCTGCGGCCGCGCTGCGCCGGGCGGCTGCCCGGACGTCTAGCTCGCGC
GGGCCGAGGCGCGGGGAGGCAGGGCGGGAGGAGGAGGGACCAGGCGGCGGAGGAGGGATCGG
GAGGAGCAGAGGAGGAGGAGAATGCGAGGAGGAGGGAGGAGAACC-3′

The notations are the same as above. Here, the reverse primer was selected – highlighted with dark green – to have the same length as the forward primer (supposed that the length of 15 base pairs upon hybridization will be specific in the above DNA sequence). Seemingly these two primers can be suitable for PCR. However, since both primers are present in the same reaction mixture, both of them must hybridize to its target strand at the same adjusted annealing temperature. This requires that their melting points shall be very close to each other. Ideally, the melting points of the two primers are identical. Thus, the melting point has to be checked. This is usually carried out by estimating this temperature. Initially, the melting point (T_m) was calculated from a simple formula:

$T_m = 4 \times (yG + zC) + 2 \times (wA+xT)$, where w,x,y,z are the number of the 2'-deoxynucleotides (A,T,G,C) in the sequence, respectively.

The use of this formula is not recommended for more than 13 nt (nt = nucleotide; ref: J. Marmur and P. Doty, J. Mol. Biol., 1962, 5, 109-118). N(nt) = y + z + w + x. For longer primers the Wallace formula is suggested:

$T_m = 64.9 + 41 \times (yG + zC - 16.4)/(wA + xT + yG + zC)$

(ref: R.B. Wallace et al., Nucleic Acids Res., 1979, 6, 3543-3557).

Nowadays on-line calculators are used for T_m estimations. Several calculators are available, examples are e.g, the Oligo-Calc (http://biotools.nubic.northwestern.edu/OligoCalc.html) and the Oligo Calculator https://www.bioinformatics.org/JaMBW/3/1/9/index.html. Using the second program, the following characteristics of the two selected primers can be obtained.

- Forward primer; 5'-CCCCGGCGCAGCGCG-3':

nt = 15; T_m = 58 °C; GC content = 93%

- Reverse primer; 5'-TTTCTTTTCCTCCAG-3':

nt = 15; T_m = 36 °C; GC content = 40%

The results of the calculations show unfortunately, that the melting points of the two primers differ significantly. The T_m of the reverse primer is much lower than that of the forward primer. In this case there is only one opportunity to increase the melting point of the reverse primer: it has to be prolonged towards its 3' end until the desired melting temperature is not attained. This procedure results the following reverse primer, in which the newly added nucleotides are highlighted with a light green background.

- Reverse primer; 5'-TTTCTTTTCCTCCAGAGCCCGAC-3':

nt = 23; T_m = 57 °C; GC content = 52%

The inverse sequence of this primer is also highlighted in the chain A above:

5'-GTCGGGCTCTGGAGGAAAAGAAA-3'.

The two primers have now their melting points very close to each other. Nevertheless, they are not ideal, since do not fulfill several other expectations against the ideal primers. These expectations are listed below as general guidelines and advices of the primer design:

- The melting temperature of the primers (T_m) should be 50-60°C.
- As it is described above, the T_m values of the primers must not differ significantly.
- Avoid long stretches of identical nucleotides in particular GGGGs
- The length of the primers should be between 18-30 bp.
- The primer pairs should not differ significantly in lengths.
- From the above it becomes clear the GC content of a primer shall be between 40% and 60%.
- while the 5' tails do not significantly affect annealing, the 3' end possibly should be G or C (stronger bond).
- Palindromic and inverted repeat sequences should be avoided.
- Complementarity between the pairs of primers will result in primer-dimers.
- The longer is a primer, the more expensive is its production. The primers longer than 30 bp require further purification, increasing the price of the primer.

- The sequence of the primer shall be written correctly in 5' → 3' direction (left to right), when placing the order of the primers.

These guidelines center upon a few important points. These are the melting, the hybridization properties of the sequence, and the economic data. As mentioned, the increase in the length of the primer increases its price. But this is not the only disadvantage of the long primers. It also has to be mentioned that the solid phase synthesis of the oligos is not a 100% accurate procedure. This means that in each step of the synthesis some of the oligonucleotides are not extended. Therefore, in the obtained product few primers of erroneous sequence will appear. The longer is the primer, the higher is the probability of the occurrence of these erroneous sequences. Although these can be altogether less than 1%, there are some applications which are prone to select erroneous genes amplified in PCR leading to wrong experiments. This is e.g, the case when the selected gene is toxic to the bacterial cell used in later experiments.

The above discussion clearly demonstrated that the primer design is not an easy task, and it requires concentration not to make mistakes. In spite of very careful work, it happens that the PCR is not successful. In this case, the experiment shall be repeated by varying the annealing temperature. It is an easy experiment with a gradient PCR, which can vary the temperature in the adjacent test tubes. Otherwise, it is a time-consuming procedure. However, new primers have to designed if this procedure fails.

The molecular weight of a synthetic 2'-deoxyoligonucleotide is calculated by the following equation:

$$M_w = (A \times 313.2) + (C \times 289.2) + (G \times 329.2) + (T \times 304.2) - 62 \text{ Da}$$

Here A, C, G and T refer to the number of the appropriate 2'-deoxynucleotides. Note that the multiplication factors are the molecular weights of these dNMPs with the molecular weight of a water molecule subtracted. The finally subtracted 62 Da shows that as the consequence of the solid phase oligonucleotide synthesis the final 5' phosphate group is removed from the molecule, which is of great significance from the point of view of the reactivity of these termini.

It is also worth mentioning that the presence of Mg^{2+} ions in optimized concentration is needed for the proper working of the polymerase enzymes. Therefore, one of the ways to introduce random mutations is to disturb the optimal amount and composition of the available metal ions in the PCR reaction mixture, and thereby the optimal function of the polymerase enzyme – when this is the goal of the experiment.

It is a great potential of the primers that they can also be applied for modification of the DNA sequence in the amplification procedure. There are several types of the modification for different purpose. One of the most common type of modification is the fusion of a short oligonucleotide sequence at the 5' termini of the primers. These short sequences are shown in **Fig. 27.** as pink/violet boxes. This can easily be understood, knowing that usually the subsequent processing of the amplified DNA leads to its insertion into a DNA carrier (DNA vector) for cloning in bacteria. To assure that this process is specific, so-called restriction endonucleases (explained in detail in a following chapter) are used to cleave the termini of the PCR product. In this way specific sticky ends are created,

which can be hybridized to another DNA cleaved with the same enzyme. The restriction endonucleases recognize few base-pair long sequences. These restriction enzyme cleavage sites can be introduced into the PCR product by the appropriately designed primers.

With a careful design of the primers it is also possible to introduce mutations into the newly amplified product. This mutation is shown in **Fig. 27.** by a yellow circle within the primer sequence. The new DNA strands will contain the modifications introduced by the primers. These will result in mutation close to the termini of the new DNA fragments. If the required mutation site is located far from the termini, a slightly different strategy can be used.

The sequence can be divided into two parts that overlap around the mutation site. The two DNA segments can be amplified in separate PCR reaction mixtures (1st and 2nd PCR in **Fig. 28**. Then, they are annealed together to give the whole desired sequence. This product, encoding the mutant protein, is serving as the template in a third PCR using the two terminal primers for its amplification. The procedure is presented schematically in **Fig. 28.** Note that the mutation site here can not be very close to the 5' end of the primers, since is this case there would be no overlap between the two fragments formed in the 1^{st} and 2^{nd} PCR, and therefore these could not be stably hybridized together.

It also has to be recognized that the primers, which do not hybridize through their full sequence, either because of the 5' prolongation or because of the mutation site inside the sequence, have different melting points from the calculated one for the full length sequence. In addition, when the mutation is already introduced in the first cycle of PCR, the new DNA strands also behave as templates in the next cycles, offering the hybridization through the full length of

the primers. Thus, the melting points of the primers change in this step of the PCR. This has to be kept in mind during the primer design.

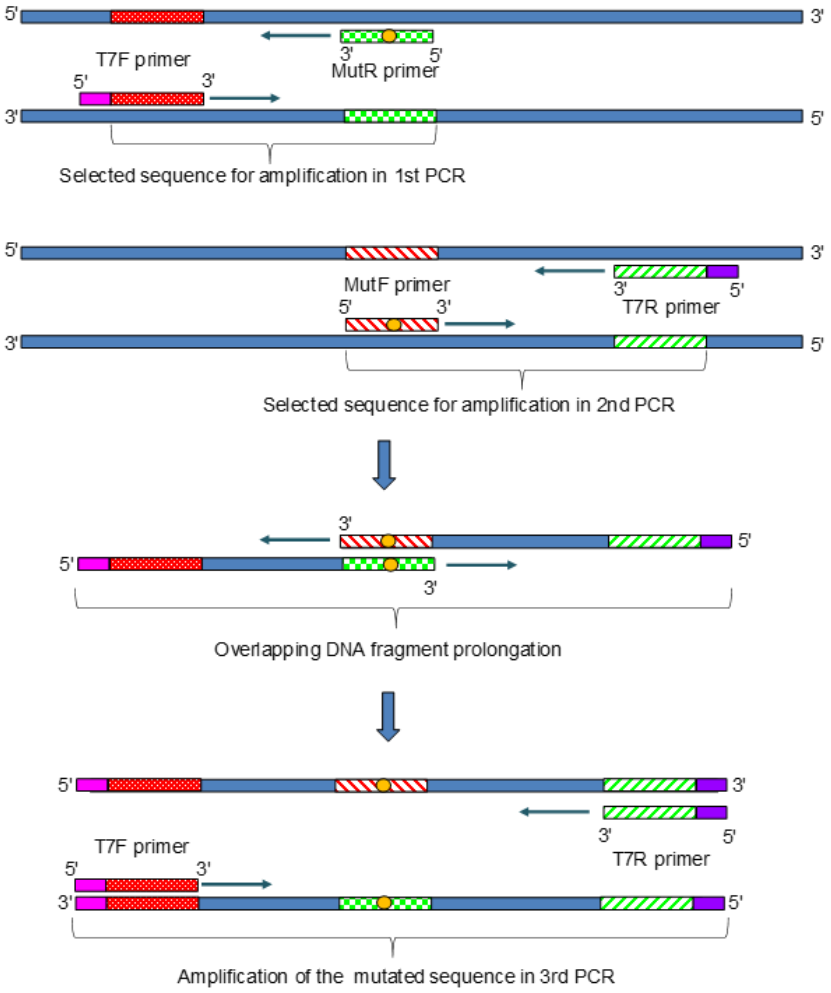


Figure 28. The schematic representation of the introduction of a point mutation by the help of designed PCR primers. The 1st and 2nd PCR are carried out in parallel (note that the melting point of all the primers shall be the same in these reactions). Then, the products of the two PCR are mixed together, and the appropriate strands are hybridized and prolonged by the DNA polymerase enzyme. The double strand DNA formed in this process will serve as the template of the third PCR, in which the full selected DNA sequence (a gene), including the designed mutation, is amplified.

Another efficient way of the mutation of DNA in order to replace amino acids, delete or insert single or multiple adjacent amino acids is the QuikChange™ method. This can be applied for genes which are already inserted in a DNA carrier. These DNA molecules are often called vectors. Their most prevalent representatives are the so called plasmids, which are extrachromosomal DNA molecules of few thousands of base-pairs within a cell, physically separated from chromosomal DNA. They are most commonly found as circular, double strand DNA molecules in bacteria, which can replicate independently, therefore they are used in DNA cloning and protein expression (see later). Such circular DNA molecules can be used as templates for PCR amplification using overlapping forward and reverse primers. Both of these primers contain the designed mutation, as shown in **Fig. 29.**

Since in this procedure the full plasmid is amplified, a high-fidelity DNA polymerase has to be applied in the PCR, which can amplify a long DNA without introducing random erroneous nucleotides in the new DNA strands. A failure in the early cycles of PCR would result in a large fraction of new DNA with unwanted modifications. It is advised to increase the amount of dNTPs in such PCR, which is needed for synthesis of the long DNA strands. This will also require the optimization of the Mg^{2+} concentration in the reaction mixture, which is necessary for the polymerase enzyme to work properly. This is usually already taken into account by the special buffers supplied together with the polymerase enzymes able to build up long DNA chains.

The use of a low amount of template is important in this procedure, because it remains together with the product as an impurity, interfering with subsequent applications (see later). Nevertheless, the original templates purified from

Escherichia coli (*E. coli*) bacteria can be degraded by a specific nuclease, which can digest the methylated template, but this causes extra cost.

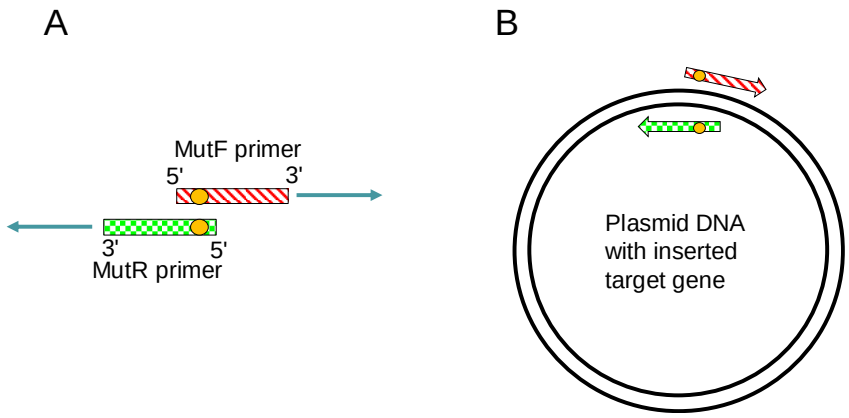


Figure 29. A) The overlapping primers including the designed mutation used in QuikChange™ method to amplify the new DNA molecule. **B**) Schematic representation of the hybridization of the primers to a double strand circular plasmid DNA.

Finally, one more challenging application of the primers will be mentioned. Imagine a research project in which a new gene is created, which can not be found in the nature. This is the situation when e.g. a new artificial enzyme is designed, as shown in **Fig. 30**. The artificial nucleases depicted in the figure are constructed of DNA fragments, which encode for an HNH catalytic domain, a ZF zinc finger protein and a control domain, regulating the properties and the function of the enzyme. These proteins were designed using a computer program and therefore, it is essential to obtain the precise amino acid sequence for experimental investigations. Thus, the gene of these enzymes has to be precisely fused together from the individual genes of the HNH, ZF and control units.

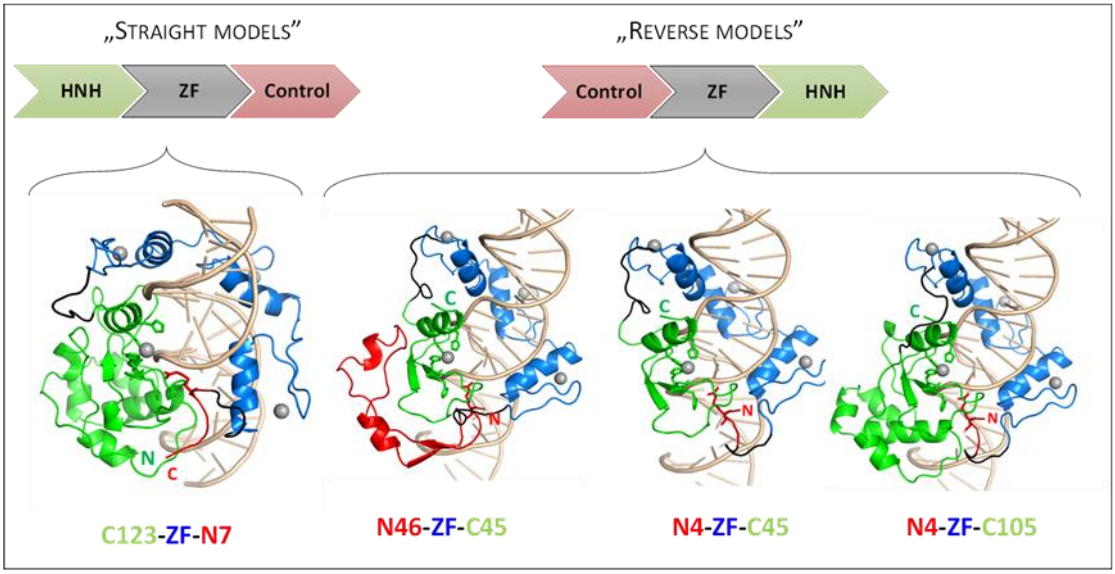
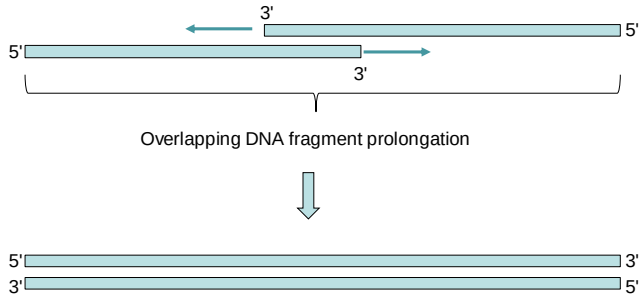


Figure 30. Artificial nucleases have been designed using a computer program suite by Eszter Németh in the frame of her PhD dissertation, conducted in the laboratory of the author of this e-book.

The principle of this procedure is the prolongation of overlapping DNA fragments as shown in **Fig. 31A**. These DNA fragments can eventually be constructed starting out from primers in a way depicted in **Fig. 31 B**.

Here the ZF unit consisting of three "finger" units is synthesized from 2'-deoxyoligonucleotides. These primers hybridize to each other by their C-termini. In subsequent DNA prolongation steps the DNA fragment size increases gradually. In the final step, the required full size DNA is obtained, but the amount of this DNA will be very small. Using the two terminal primers (ZnN1 forward and ZnC reverse primer), however, it is possible to amplify this DNA fragment in a PCR.

A



B

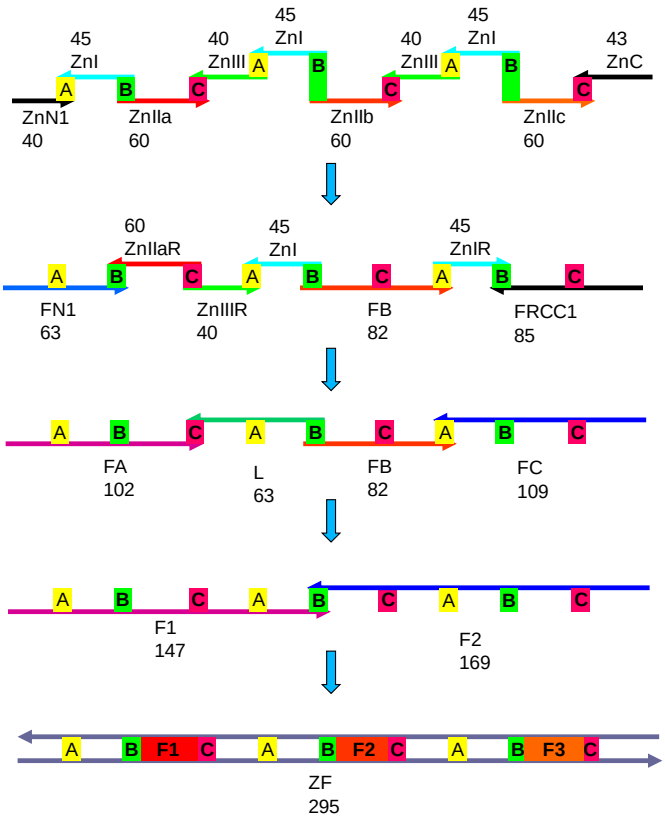


Figure 31. A) Prolongation of overlapping DNA fragments results in a new double strand DNA the termini of which are determined the 5' termini of the two original DNA fragments. **B**) Design of the primers for ZF construction. A, B and C represent the overlaps, while the numbers show the size of the DNA fragments.

As there is no template in this procedure, the DNA codes of amino acids can be chosen by the researcher (note that some amino acids have multiple codons – see later). First the so called consensus sequence is constructed, containing unusual 2'-deoxyoligonucleotide codes as listed in **Table 1**, and then the appropriate 2'-deoxyoligonucleotides are selected. The advantage of this is that the sequence of the primers can be designed in an optimal way, i.e. 50% of GC content can easily be achieved.

As an example, the following primer sequence is obtained in the first step:

5'–TGYCCNGARTGYGGNAAR-3'

The A or T dNMPs are highlighted with yellow (total 5), while the G and C with blue (total 7) background. The primer length is 18 nt thus, ideally 9 A or T and 9 G or C has to be included. The choice is multiple, since Y = C or T, R = A or G and N can be A or T or G or C. Using any of these will result in the same amino acid sequence. As an exercise write several primer sequences based on the above written one with the same length and 50% GC content.

Monitoring questions

- What are the primers used for in PCR?

- What are the basic rules of the primer design?

- Why it is necessary to design the primer pairs in a PCR to have the same melting points?

- How are the 2'-deoxyoligonucleotides synthesized, and what are the consequences of this?

- Describe the principle of introducing a point mutation into a PCR product!

- What are plasmids?

- Describe the principle of introducing a point mutation into a plasmid in easiest way!

- How it is possible to construct a new artificial gene using 2'-deoxyoligonucleotides?

6. Identification of PCR products – agarose gel electrophoresis

Students who study this chapter will acquire the following specified learning outcomes:

Knowledge

The students understand the basics of gel electrophoresis

The students know the role of the solid support in electrophoresis.

The students list the requirements towards the solid support.

The students understand the sample preparation steps for the agarose gel electrophoresis.

The students list the factors influencing the migration of the DNA in the gel.

Skills

They students select appropriate molecular weight markers.

The students identify DNA fragments visualized in the solid support by molecular weight marker.

The students decide about the properties of the agarose gel for optimal performance in DNA fragment separation.

The students evaluate the results of the agarose gel electrophoresis experiment.

The students cast agarose gel.

Attitude

The students take effort to design a successful agarose gel electrophoresis experiment.

The students take care of adjusting the proper conditions for the agarose gel electrophoresis experiment.

The students document the gel electrophoresis experiment and accurately process the gel photo obtained.

Responsibility and autonomy

The students design their agarose gel electrophoresis experiments independently, taking into account the various factors that influence the mobility of the DNA fragments.

The students make effort to explore the further possibilities of the agarose gel electrophoresis in DNA investigations.

To verify the success of PCR, the products are visualized by means of electrophoresis as it was already demonstrated in **Fig. 26**. This procedure is based on the different mobility of ions in solution upon applying electric field. In aqueous solution itself, the difference in the migration ability of various ions is negligible and in addition, the increase of the temperature increases the flow of the solvent. Therefore, a solid matrix is applied to separate the ions. The 3D network of the pores within the solid support decreases the diffusion compared to the liquid and also enhances the separation efficiency.

However, the solid supports must fulfill several requirements to be applicable in electrophoretic systems:

(i) They should be hydrophilic, as the electrophoresis is carried out in aqueous buffer solution.

(ii) Chemical inertness is necessary. The solid support should not react with the substances applied for separation.

(iii) The solid support shall be neutral i.e. with no charges, which could interfere with the migration of the ions.

(iv) Adjustable pore size is an advantageous property of the solid support, as it provides flexibility to carry out the separation of ions with various properties.

(v) The solid support shall be physically rigid, it has e.g. to endure the transport form the electrophoresis tank to the documentation chamber.

(vi) It has also to be transparent to guarantee the visibility of the various dyes applied in electrophoresis.

(vii) It should not absorb pigments which are applied in electrophoresis.

The most abundant solid supports are gels made from agarose or polyacrylamide. As the former is applied more often for the investigation of DNA, the polyacrylamide gels will be discussed in the chapter dealing with proteins later. Agarose is a polysaccharide, purified from agar or agar-bearing marine algae. It is a linear polymer of the repeating unit of agarobiose. The building block (monomer) of the agarose is depicted in **Fig. 32**.

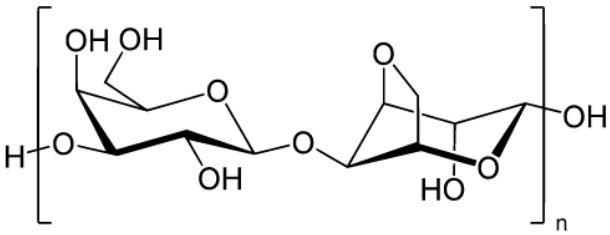


Figure 32. Agarobiose (4-O-*β*-D-galactopyranosyl-3,6-anhydro-L-galactose) is the monomeric unit of the agarose consisting of D-galactose and 3,6-anhydro-L-galactose units

It is a white solid powder, to be weighted on a balance for gel preparation. It is then transferred into a buffer or water and heated up usually in a microwave oven until the solution becomes clear and homogeneous. In certain laboratories the buffer used in gel electrophoresis is added to this solution when it cools down to ~ 50°C in its 50 × concentrated form, in a way that the final concentration of the buffer should be the required one according to the protocol. Commonly TAE

of TBE buffers are used in agarose gel electrophoresis. The composition of the concentrated buffers is shown in **Table 2**.

As an easy exercise, calculate the final molar concentration of the constituents (for molecular weights and densities refer to external sources). Also check the price of 500 mL 50 × concentrated buffers, described in the table, based on the information available on the web in their ready form, and put together by the researcher.

Table 2. The composition of the concentrated buffers used for agarose gel preparation.

50 × concentrated TAE	**5 × concentrated TBE**
121 g of Tris base	54 g of Tris base
28.5 mL of acetic acid	27.5 g of boric acid
50 mL of 0.5 M EDTA	20 mL of 0.5 M EDTA
water up to 500 mL	water up to 1 L

After mixing of the solution, the gel casting follows. The solution is transferred into a casting tank. A comb is placed in the cast to create wells for sample loading, as shown in **Fig. 33**. The liquid is then allowed to cool down below the gelling temperature. This temperature is dependent on the source of the agarose, usually it is in the 35–42 °C range. The three-dimensional matrix of the agarose gel is formed of the agarose polymers in supercoiled bundles. The gel is usually completely set within ~ 30 minutes.

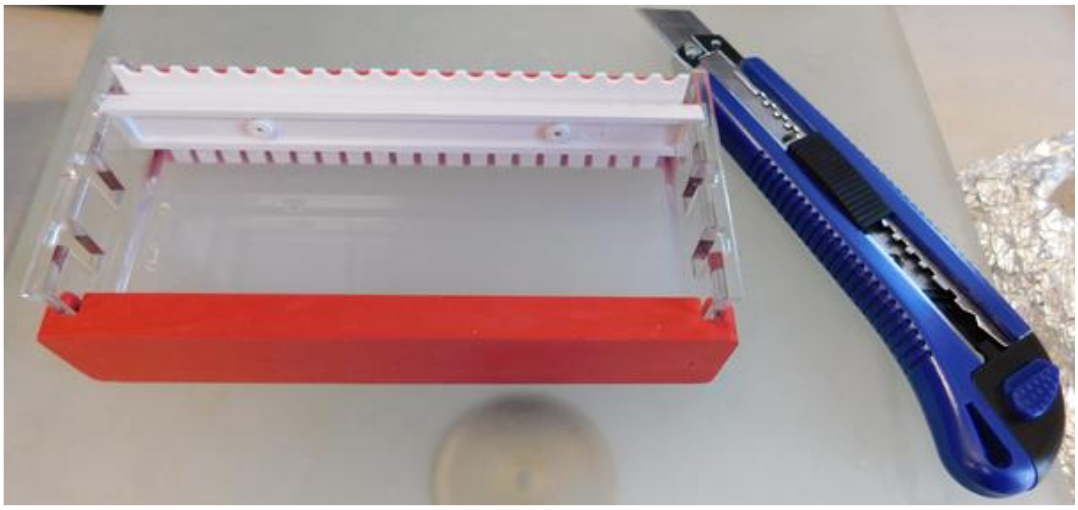


Figure 33. Agarose gel casting. The comb has to be set properly to create the wells for sample loading. (Take care not to punch out the gel with the comb when it is set and when it is removed.) The scalper is used to cut the gel into pieces required for the gel electrophoretic experiment.

Pores and channels are formed in the three dimensional structure of the gel formed in this process, through which charged biomolecules can migrate. For investigation of DNA fragments, the most frequently applied technique is the agarose gel electrophoresis (AGE). The negatively charged DNA-molecules migrate towards the positive electrode. They are separated by their size in agarose gel: the shorter DNA fragments (lower molecular mass) migrate more quickly, while the longer ones more slowly. Naturally, the size of the pores and channels will determine the migration ability of the analytes. During the design of the agarose gel electrophoresis experiment both the pore size of the gel and the expected size of the DNA molecules to be separated shall be taken into account. The larger is the DNA molecule the more effectively the matrix impedes its migrations. However, too small pores and channels will prevent the large DNA

molecules to be separated from each other, as they all migrate very slowly. In the contrary, the gel with an increased pore size will not be suitable for the separation of small size DNA fragments, as they all will migrate with the same maximal velocity. As a general guideline, the advised concentrations of agarose gels for the separation of DNA molecules of certain size ranges are provided in **Table 3**.

Table 3. The suggested concentration of agarose for separation of various DNA mixtures in AGE.

Agarose concentration (g / 100 mL)	**Optimal DNA resolution (kbp)**
0.5	1 – 30
0.7	0.8 – 12
1.0	0.5 – 10
1.2	0.4 – 7
1.5	0.2 – 3

As an exercise, make a decision about the concentration of the gel for ideal separation of the PCR products shown in **Fig. 26** (note that the DNA marker is needed for the identification of the size of the fragments – see later).

The set agarose gel is transferred into the horizontal electrophoresis tank, and it is immersed into the buffer solution with 1 × concentration. The level of the buffer solution shall be high enough to fully cover the gel. Make sure that the gel is placed in the right direction, i.e., the wells shall be close to the negative electrode, as the DNA molecules will migrate toward the positive electrode. In the next step the samples are loaded into the wells of the gel using digital

Finnpipettes for liquid handling (a wide range of such pipettes are on the market), as it is shown in **Fig. 34**.

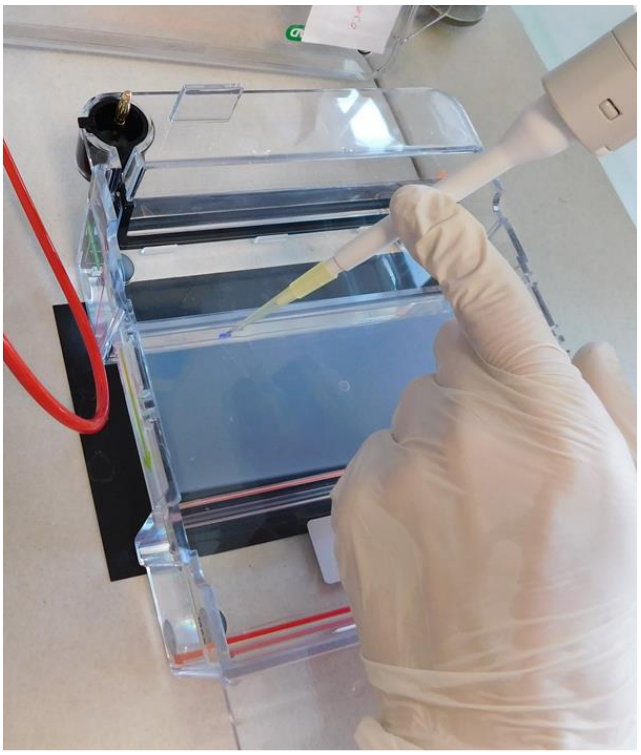
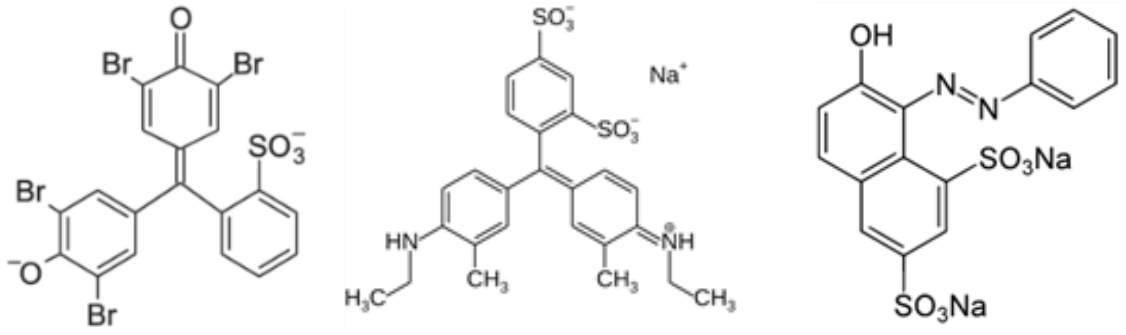


Figure 34. Sample loading in for agarose gel electrophoresis. The volume of the sample is usually few μL. The handling of such small volumes is done by precision digital Finnpipettes. Plastic tips are attached to the pipettes, which have to be replaced for each new sample. By means of this, the contamination of the original and the loaded samples can be avoided. (Notice the blue color of the DNA sample.)

DNA is well soluble in aqueous buffer solutions. This makes it difficult to load into the wells and prevent its immediate diffusion into the electrophoresis tank. To avoid this the density of the DNA solution has to be increased. For this purpose, glycerol or concentrated sucrose solution is used. Such materials are usually used in the "loading dye" solutions optimized for DNA sample loading

and visualization. The latter is important not only for safe sample loading but also to visually follow the migration of the DNA during AGE. Thus, various dyes (**Fig. 35.**) are also mixed into the loading dye solution, such as Bromophenol Blue, Xylene Cyanol, Orange G, etc.



Bromophenol Blue **Xylene Cyanol** **Orange G**

Figure 35. The most common dyes applied in AGE for the visualization of DNA loading and migration.

The DNA sample solution is mixed with the 6 × or 10 × concentrated loading dye solutions prior to loading it into the wells of the agarose gel. As it can immediately be noticed from **Fig. 35**, all the applied dyes are negatively charged under the conditions of the AGE, thus, they will migrate in the same direction as the DNA. The different dyes migrate with a different velocity depending on the gel concentration. Nevertheless, apart from examining the behavior of the dyes in the specific setup, general experience may help conducting an AGE experiment. According to the observations, **Table 4.** shows the relationship between the migration of the DNA and the dyes.

Table 4. Observations on the relationship of the migration of DNA and pigments commonly applied in loading dyes during AGE.

Dye	**0.5–1.5 g / 100 mL agarose gel**	**2.0–3.0 g / 100 mL agarose gel**
Xylene Cyanol	4–10 kbp	0.2–0.8 kbp
Bromophenol Blue	0.4–0.5 kbp	< 0.15 kbp
Orange G	< 0.1 kbp	n.d.

When all the samples are properly loaded, the gel electrophoresis is initiated by applying a high voltage between the anode and cathode. For a common experiment, a 7 V / cm potential gradient factor is advised, which has to be multiplied by the distance between the two electrodes (in cm) for the proper potential. The start of the electrophoresis is accompanied by visible evolution of gas bubbles at the electrodes, as the electrolysis of the water occurs as a side-reaction. The electrophoresis can usually be terminated after 20–30 minutes, but the process also can be visually followed through the migration of the dye additives. A running AGE experiment is shown in **Fig. 36**. Because of the applied high voltage, it is very dangerous to touch the gel during the process. To prevent this, many electrophoretic instruments are already designed in a way that the electric circuit is disconnected by removing the cover from the electrophoresis tank – for safe operation.

As an exercise, try to identify the dyes on the gel and decide, whehter the experiment has to be stopped now if you would like to separate DNA fragments

in 3.0 – 6.0 kDa size range, supposed that 1 g / 100 mL concentration agarose gel (often denoted as 1 % in the literature) was used here.

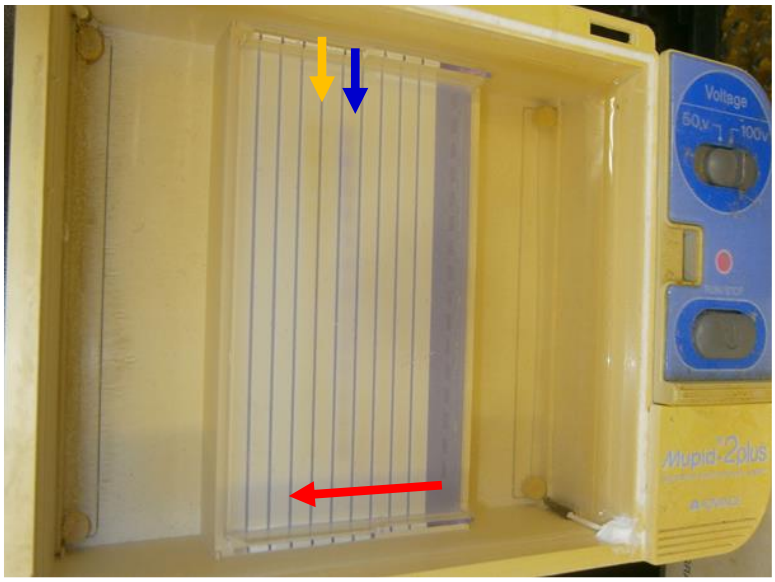


Figure 36. The agarose gel in the course of AGE in a simple instrument. The red arrow indicates the direction of the migration of the DNA and the negatively charged dyes, while the actual position of the two applied dyes in the loading dye are shown by the blue and yellow arrows.

A care has to be taken of the precise conditions of the AGE experiment, since the buffer in the tank is warming up during the procedure. As the structure of the gel is held together with hydrogen bonds, it can be disrupted by warming up to its melting temperature close to 85–95 °C. In this case, the experiment will be unsuccessful. But warming up to ~50 °C will also soften the gel, so that it will be difficult to transfer it to the documentation phase, as well as the migration of the DNA is also altered in warm buffer. If necessary, the buffer has to be cooled.

After finishing the electrophoresis the DNA molecules have to be visualized in the gel. This is most often achieved by reacting the DNA with an intercalating reagent, ethidium bromide. Because of this property of ethidium ion, it is carcinogenic – thus, wearing gloves when working with it is a must!!! There are newly developed dyes for detecting DNA, such as GelRedTM, which is claimed to be safe to use, without carcinogenic effects. The ethidium ion becomes fluorescent, emitting orange light with a wavelength maximum of 605 nm when intercalated into the DNA double strand and irradiated with UV light (**Fig. 37.**). The fluorescence of the dye itself is negligible in comparison to the intercalated one, but it causes a slight orange background.

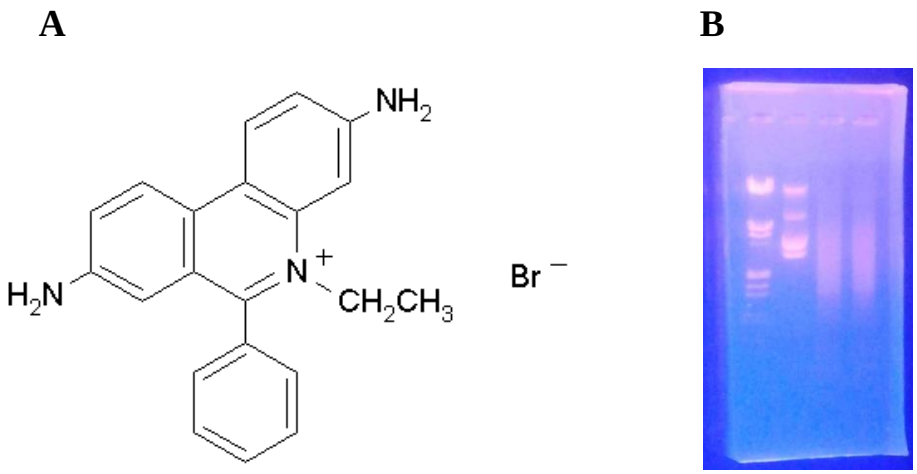


Figure 37. A) The structural formula of the ethidium bromide DNA intercalating agent. **B**) An agarose gel with DNA visualized by ethidium bromide.

The reagent can be supplied by immersing the ready gel into an ethidium-bromide solution, but it also can be added directly to the agarose in the same step of the preparation of the gel, as it was applied for the 50 × buffer. Ethidium bromide may be decomposed at high temperatures, therefore, it should be only

added when the agarose solution cools down to ~50 °C and then the solution has to mixed well to achieve homogeneous distribution of the dye. This latter method has several disadvantages. The main is the carcinogenicity of the reagent, which requires special handling of all the AGE reagents and instruments with gloves. It also has to mentioned that the ethidium ion is positively charged, thus it migrates in the opposite direction in the gel as compared to the DNA. This is well recognized in the gel shown in **Fig. 37B**. The background fluorescence in the lower part of the gel has disappeared, as there is no ethidium bromide. This phenomenon makes it more difficult to identify the small size (< 200 bp) DNA fragments if they run close to the bottom of the gel. In addition, the small size DNA can bind less fluorescent molecules, which further decreases its fluorescence. The ethidium ion intercalated to the DNA may also influence the migration behavior of the DNA fragment. On the other hand, staining the gel by immersing it into ethidium bromide solution for ~ 30 min is not as efficient as the direct involvement of the dye in the gel, and also it promotes slight diffusion of the DNA in the gel resulting in the decrease of the picture definition. The optimal solution has to be chosen for the relevant experiment. The result of every experiment has to be documented. For this purpose, in AGE several advanced gel documentation instruments have been developed. The one used in the laboratory of the author is depicted in **Fig. 38.** This instrument consists of a UV transilluminator and a digital camera system. The photos can be transferred to computers for further processing.

A **B**

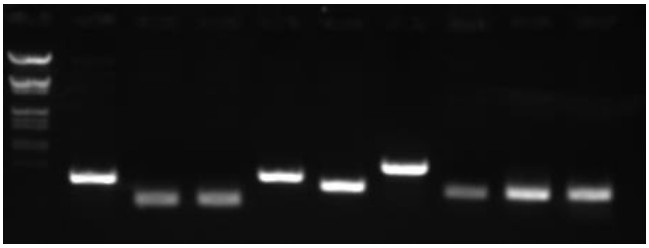


Figure 38. A) An AGE gel documentation system **B**) An agarose gel documented, which needs always to be processed and labelled properly, e.g. as it is shown in Fig. 26.

The multiple bands usually in the first lane of the gel arise from a DNA mixture of known size DNA fragments. This is called DNA marker (or DNA ladder), based on which the size of the investigated DNA sample is estimated. Running this marker only in a parallel with the sample in the same gel, the size of the unknown DNA fragment can be obtained properly, since the migration of the DNA molecules in the gel depend on various factors as mentioned above. Several DNA markers are available on market containing different ranges of sizes of the DNA fragments, as shown in **Fig. 39.**

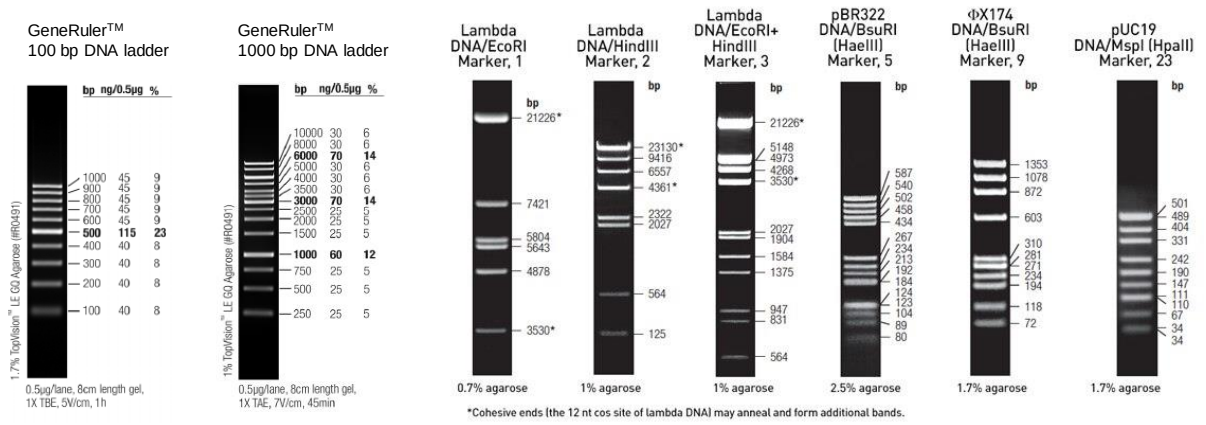


Figure 39. Various markers from various suppliers can be chosen for the AGE experiments. Select always the appropriate one for identification of the investigated DNA.

It is worth mentioning that the migration of the DNA molecules in the agarose gel also depends on the shape of the DNA. Most of the markers consist of only linear DNA molecules. Thus, the bands of the marker aro only comparable with the bands of linear DNA. This is appropriate for detection of the PCR fragments. However, it was already mentioned, that plasmids are circular DNA molecules recognized by bacteria. These circular molecules can exist in several condensed supercoiled forms, topological isomers. Topoisomerase enzymes are able to interconvert between such topological states. A plasmid preparation usually contains superhelical (or supercoiled) form of the DNA, but depending on the skills of the researcher, and the protocol applied, the so called open circular (relaxed) form of the plasmid DNA also appears (**Fig. 40.**). The amount of the latter form can also be increased by introducing a single strand cleavage (nick) in the superhelical DNA.

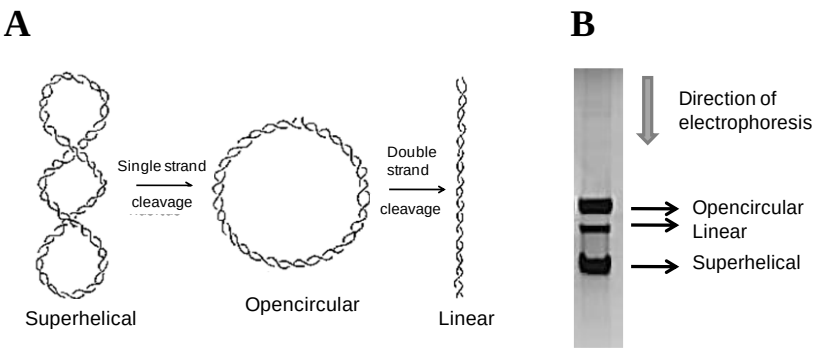


Figure 40. A) Different forms of bacterial plasmid DNA which can be converted into each other by the help or various enzymes. **B**) The result of the AGE experiment carried out with a plasmid DNA sample.

The circular DNA can also be linearized by introducing a double strand cleavage. Looking at **Fig. 40.** it can be easily suggested that the most compact superhelical form migrates most quickly in the agarose gel, while the bulky open circular form is usually very slow. The band of the linear form is usually detected between the two above forms.

Monitoring questions

- Which method is suitable for easy verification of the success of PCR?

- How can be the various ions distinguished by an electrophoresis experiment?

- What is the most commonly used solid support in the electrophoretic separation of DNA mixtures?

- What is the basic principle of the separation of DNA molecules by agarose gel electrophoresis?

- What kind of agarose gel shall we prepare for the separation of a mixture containing large, and what kind for the separation of a DNA mixture containing small DNA molecules?

- What is the composition of the "loading dye"? What is the role of the pigments in the "loading dye"?

- How can the DNA molecules be detected in the agarose gel? Which important precautions have to be considered during this procedure?

- How can the DNA fragments be identified in the gel picture?

- Which properties of the DNA influence its migration in the agarose gel?

- Which environmental factors influence the agarose gel electrophoresis experiment?

7. Restriction endonucleases and DNA cloning

Students who study this chapter will acquire the following specified learning outcomes:

Knowledge

The students explain the concepts of DNA cloning.

- The students know how the double strand DNA is cleaved by restriction endonucleases.

- The students understand the importance of the restriction/methylation system

- The students list the guidelines of the selection for restriction site(s) in a DNA cloning experiment.

- The students are aware of the factors influencing the efficiency of the cleavage of the PCR fragments by restriction endonucleases.

The students know how the amount of the enzyme is provided by the supplier.

The students explain the meaning of the cloning vector, and list various types of the cloning vectors.

The students are aware of the meaning of the multiple cloning site.

The students understand the plasmid DNA construction.

Skills

The students select restriction site(s) for DNA cloning experiment based on various requirements.

The students identify palindromic sequences as restriction sites within DNA fragments of various lengths.

The students recognize the advantage of using two different restriction enzymes for DNA cloning. They consider also the disadvantages.

The students calculate the statistical abundance of a restriction site within a large genomic DNA.

The students determine the concentration and the purity of the DNA solution by spectrophotometric method.

The students distinguish the transformed and non-transformed bacterial cells.

The students distinguish the cells transformed by the ligated and non-ligated plasmids.

Attitude

The students pay attention to the importance of correct design of the oligonucleotide primers, including restriction sites.

The students take care of the proper handling of hazardous materials.

The students are critical while evaluating the gel electrophoresis of plasmid DNA.

Responsibility and autonomy

The students realize the importance of the control experiments and explain these to their colleagues.

The students phrase independent suggestions about the selection of the optimal strategy for DNA cloning.

The DNA fragments encoding for proteins, i.e. the genes amplified by the PCR have to be built in a suitable carrier DNA, so called plasmid, to use them for expression of proteins within the bacterial cells. The summary of this procedure is shown in **Fig. 41**.

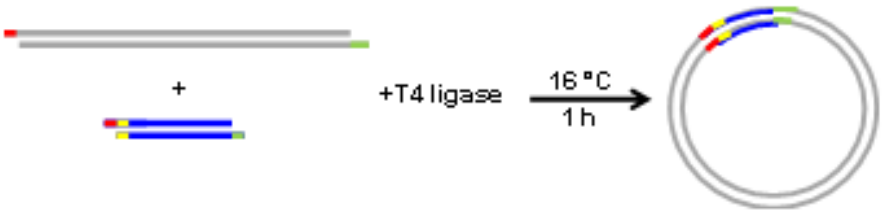


Figure 41. Both the plasmid and the PCR fragment are prepared by configuring their termini for their fusion, which is then performed with the help of an enzyme called ligase.

The discovery of so called restriction endonucleases was a crucial milestone in this procedure. The term restriction enzyme is originated from the investigations of bacteriophage λ, a bacterial virus and the phenomenon of the bacterial host-controlled restriction and modification. This biological process was first identified in work done in the laboratories of Salvador Luria and Giuseppe Bertani in the early 1950s. The observations showed that bacteriophage λ can grow well in one strain of *E. coli*, but the growth in another strain was significantly restricted. The latter host cell has the ability to reduce the biological activity of the bacteriophage λ.

In the 1960s, the experiments carried out in the laboratories of Werner Arber and Matthew Meselson revealed that the restriction is caused by enzymatic cleavage of the phage DNA, caused by an enzyme denoted as restriction enzyme. Inside a prokaryote, the selective cleavage of the foreign DNA prevents viral infection. On the other hand, the host DNA should be protected against the DNA hydrolysis. This is attained by the selective methylation of the host DNA with a methyltransferase enzyme. Such modification of the prokaryotic DNA inhibits the hydrolytic cleavage. These two enzymes together form the restriction/methylation modification (RM) systems of bacteria. A restriction/methylation system cuts the DNA chain or methylates selected DNA bases at or near specific recognition nucleotide sequences. These sequences are called restriction sites. The first restriction endonuclease enzyme was isolated and identified by Hamilton Smith and Daniel Nathans at the end of 1960s.

Restriction endonucleases are categorized into four groups:

-Type I enzymes are multifunctional. They possess both hydrolyase and methylase activities. The cleavage occurs outside the recognized DNA sequence. These enzymes require ATP and S-adenosyl-L-methionine cofactors for their function.

-Type II enzymes perform only the hydrolytic cleavage of the DNA, while the methylation is carried out by an independent enzyme. The two enzymes are encoded by separate genes. The cleavage occurs within the recognition site or very close to it. Most of these enzymes are metalloenzymes, working mostly with Mg^{2+}-ions. Some may contain even more metal ions. One example is presented in **Fig. 42**. Therefore, these enzymes are also of interest to bioinorganic chemists. It is known that metal ions participate in the reaction at various levels: they bind and

electrostatically activate the phosphodiester bond, generate a nucleophilic OH^- in collaboration with the amino acid side chains of the protein, activate water molecule for protonation of the leaving alcoholate group. Presumably because of their versatility, mostly the Mg^{2+}, Ca^{2+} and Zn^{2+} non-redox active metal ions have been chosen as catalysts during the evolution. The essential role of the metal ions also draws the attention of the researchers, that strong chelating agents, such as EDTA may interfere with the enzyme function, it may inhibit the enzyme by removing the metal ion from the active centre.

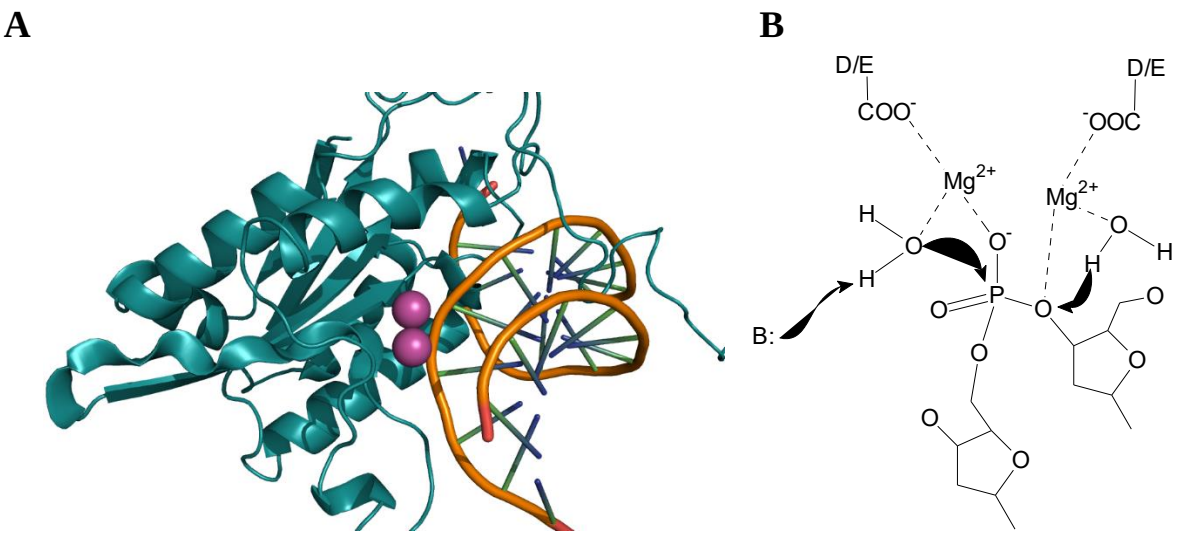


Figure 42. A) PyMol image of a BamHI restriction endonuclease monomer bound to DNA. The figure was constructed based on the crystal structure coordinates downloaded from RCSB Protein Databank. PDB Id: 2BAM. The two metal ions are close to the DNA strand, most probably participating directly in the catalytic process. They are highlighted by pink spheres. B) The suggested mechanism is shown schematically.

-Type III enzymes are complex enzymes including the methylase as well. They cleave within a short distance from a recognition site.

-Type IV enzymes recognize and cleave already modified e.g. methylated, hydroxymethylated and glucosyl-hydroxymethylated DNA.

Out of these, Type II restriction endonucleases are most commonly used for DNA cloning experiments. To cut the double strand DNA, restriction enzymes make two incisions, once through each sugar-phosphate backbone.

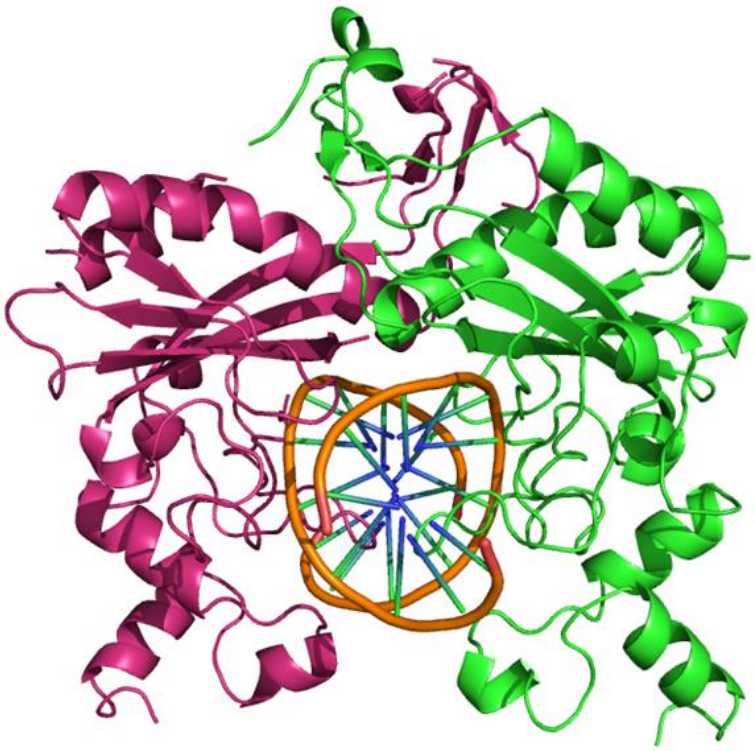


Figure 43. PyMol image of EcoRV restriction endonuclease bound to DNA in homodimeric attachment. The figure was constructed based on the crystal structure coordinates downloaded from RCSB Protein Databank. PDB Id: 1AZ0. The two protein molecules are depicted in pink and green, while the DNA is in orange.

For the efficient double strand cleavage, they work as homodimers each cleaving one strand. Both monomers bind and cleave at the recognition sequence, i.e. the sequence of the restriction site. The two monomers are centrosymmetrically arranged in their DNA bound forms, as shown in one example in **Fig. 43**.

Since both monomers are identical, they recognize the same sequence. This means that the same sequence (remember that the sequence of the DNA has the direction, which was defined from 5' to 3', and the two DNA strands are antiparallel) has to be present in both DNA strands. The consequence of this is that many of these enzymes recognize so-called palindromic DNA sequences.

Palindromic structures are also known from the grammar. These are words or phrases that read the same backward and forward. Few examples are: "Telegram, Margelet!"; "Was it a rat I saw?"; "Madam, in Eden, I'm Adam"; "Amore, Roma".

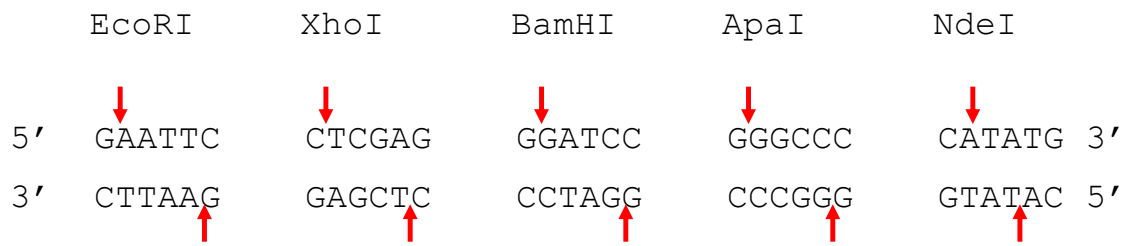


Figure 44. Palindromic recognition sequences of some restriction endonucleases. The names of the enzymes are in the first row above the restriction sites. Not that the sequences are read the same from 5' towards 3' termini. The red arrows show the site for the hydrolytic cleavage of the phosphodiester backbone.

Similarly to these, the DNA restriction sites are centrosymmetric, but on DNA terminology palindromic sequences are read the same on both strands from 5' towards 3' termini. Few such palindromic structures are listed in **Fig. 44**.

Usually, the palindromic sequences are very short. The number of base-pairs in the recognition sequence determines how often the site may appear in any given DNA sequence. A sequence with the length of n base pairs would theoretically occur once in every DNA of 4^n nt bp length. E.g., a 4 bp restriction site occurs statistically once in a $4^4 = 256$ bp sequence. The longer is the restriction site sequence the more specific is the restriction enzyme. (Statistically a 6 bp site occurs once in $4^6 = 4096$ bp, and an 8 bp site in $4^8 = 65536$ bp). At the same time these numbers also refer to the average length of the DNA fragments resulting from the restriction endonuclease treatment of a long DNA. This suggests that the genome of the bacteria will also contain some of these restriction sites. For this reason, the cleavage of its own DNA is prevented by the methylation of the DNA as mentioned above.

The cleavage site of the phosphodiester backbone is shown by red arrows for each enzyme depicted in **Fig. 44**. Note that the cleavage also occurs between the same 2'-deoxynucleotide units on both strands of the DNA. This may result in various types of cleavages, yielding 3' or 5' protruding termini or blunt-ended DNA fragments. **Fig 45.** shows the various results of the cleavage of the palindromic sequences and the termini of the cleaved fragments. It is also important to mention that the cleavage of the DNA result in 5'-P and 3'-OH termini.

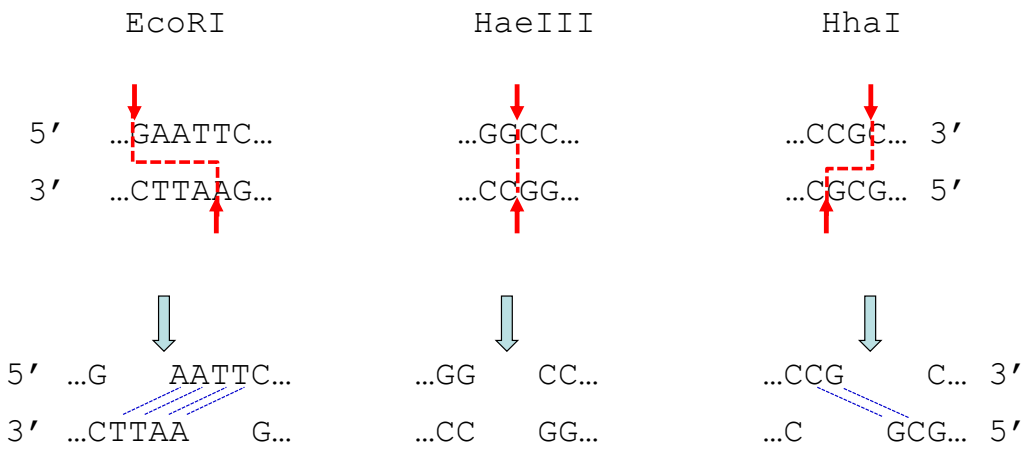


Figure 45. Examples of the cleavages with the restriction endonucleases recognizing palindromic sequences. The red lines show the pathway of the cleavage and strand separation. The blue arrows direct towards the resulting protruding or blunt termini. Note that the protruding termini can recognize each other by Watson-Crick base pair formation, so that they can hybridize and stick together the two DNA fragments having complementary protruding ends by non covalent interactions as it is indicated by the blue dotted lines.

It can easily be recognized that the protruding termini formed after the cleavage are complementary to each other. This suggests that two DNA fragments cleaved by the same restriction enzyme, can stick together through the hybridization of these termini, allowing the ligase enzyme to couple the two fragments together also by covalent bonds (see later). Thus, the potential use of these enzymes in DNA cloning became invaluable since their discovery. For these achievements, Nobel Prize has been awarded to Werner Arber, Daniel Nathans, and Hamilton O. Smith, shown in **Fig. 46**. Today more than 3000 restriction endonucleases are known, which recognize around 200 different DNA sequences in total.

Figure 46. In 1978 Nobel Prize for Physiology or Medicine was awarded to Werner Arber, Daniel Nathans, and Hamilton O. Smith (from left to right) for their work in the discovery and characterization of restriction enzymes, and their application to problems of molecular genetics. (Photo from the Nobel Foundation archive.)

For understanding the nomenclature of the restriction endonucleases it is worth mentioning that each enzyme is named after the bacterium from which it was isolated. The naming system is based on bacterial genus, species and strain. Examples of how the names of few restriction enzymes were derived are described in **Table 5**. To learn more about the restriction endonuclease cleavage sites the reader is referred to the literature and various websites. Among the latters the Restriction Enzyme Database at http://rebase.neb.com/rebase/rebase.html was often used by the author. On this website a very useful function is found among the Tools: the REBASE Tools, by means of which the investigated DNA sequence

can be searched for cleavage sites of the restriction enzymes. Based on this tool the appropriate enzymes can easily be selected for the cloning experiment.

Table 5. Nomenclature of the restriction endonucleases is presented here by few commonly used enzymes.

Name	**Bacterial strain**	**Order of identification**
BamHI	*Bacillus amyloliquefaciens H*	I
EcoRI	*Escherichia coli RY13*	I
EcoRV	*Escherichia coli RY13*	V
NdeI	*Neisseria denitrificans*	I
XhoI	*Xanthomonas holcicola*	I

The selection of the restriction endonucleases for a specific experiment depends on numerous factors. One very important guideline is that the restriction enzyme should not carry out cleavage at other sites than required. Thus, the restriction site should occur in the processed DNA molecules only at the desired cleavage position. Otherwise, the DNA molecules will be fragmented by the enzyme in an unwanted manner and spoil the experiment.

The endonuclease name also refers to another specific feature of these enzymes. The cleavage of the phosphodiester bond occurs within the DNA sequence, i.e. it is not the terminal nucleotide unit cleaved off. This would be the function of the exonucleases. This implies that the restriction site close to the

termini of the DNA fragment will be cleaved more inefficiently. Therefore, the primers introducing the restriction site sequences close to the termini of the PCR product have to be carefully designed.

From the experiments performing the cleavage of a series of short, double-strand oligonucleotides that contain the restriction endonuclease recognition sites one can obtain help for the primer design. Several tables are available containing data similar to those examples in **Table 6**.

Table 6. Efficiency of the restriction endonucleases in cleaving short double strand 2'-deoxyoligonucleotides at various times of the DNA digestion experiments.

Name	**dsDNA**	**% cleavage 2 hours**	**% cleavage 24 hours**
BamHI	CGGATCCG	10	25
	CGGGATCCCG	>90	>90
	CGCGGATCCGCG	>90	>90
EcoRI	GGAATTCC	>90	>90
	CGGAATTCCG	>90	>90
	CCGGAATTCCGG	>90	>90
NdeI	CGCCATATGGCG	0	0
	GGGTTTCATATGAAACCC	0	0
	GGAATTCCATATGGAATTCC	75	>90
	GGGAATTCCATATGGAATTCCC	75	>90
XhoI	CCTCGAGG	0	0
	CCCTCGAGGG	10	25
	CCGCTCGAGCGG	10	75

From the above data it can be concluded that the enzymes behave very differently. For the design of a BamHI restriction site at the terminus of the PCR fragment it will be enough to add two additional 2'-deoxynucleotides for the efficient cleavage with the enzyme:

5'-CGGGATCCTTAGCCGGTAAGGCCTAT-3'

However, for NdeI it is advised to add more additional 2'-deoxynucleotides:

5'-GGAATTCCATATGTTAGCCGGTAAGGCCTAT-3'

Thus in the latter case a longer primer is needed for the PCR and the subsequent cloning experiment.

Another key factor of the efficient DNA cleavage by the restriction endonucleases is the optimal buffer composition. The working buffers usually contain the following components:

-Tris-HCl (stabilization of pH)

-NaCl (adjusting the ionic strength)

-dithio-threitol (DTT) (for protection of the thiol groups from oxidation)

-$MgCl_2$ (necessary for the nuclease activity)

-Bovine serum albumin (BSA - protection against protein denaturation)

Restriction enzymes can be purchased from various suppliers. They are characterized by the unit of restriction endonuclease activity. One unit is defined as the amount of enzyme required to produce a complete digest of 1 µg of ds DNA (or fragments) in a total reaction volume of 50 µl in 60 minutes under optimal assay conditions. The enzymes are usually quite expensive and sensitive. Therefore, care has to be taken of storing them at appropriate conditions. This is usually –20 to –30 °C in a safely operating, non defrosting freezer. The solutions of the enzymes contain glycerol, preventing them to freeze under such conditions,

as the repeated freeze/thaw cycles would destroy the enzyme easily. When the enzymes are removed from the freezer for use, they have to be kept on ice, and the time outside the freezer has to be minimized.

The enzymatic reactions can be terminated by elevating the temperature or adding EDTA to bind the Mg^{2+}-ions. The protein can be removed from the reaction mixture by the general method of extraction with a phenol:chloroform:isoamyl alcohol 25:24:1 mixture saturated with 10 mM Tris, pH 8.0, 1 mM EDTA. However, this reagent is hazardous, thus the safety information provided by the supplier shall be carefully studied and the work has to be carried out accordingly.

In summary of the above, processing of the amplified genes and the DNA carrier (a plasmid) by restriction endonucleases is the initial step of the DNA cloning process. Plasmids are circular DNA molecules of bacterial origin, consisting of a few thousands of base pairs, widely applied in recombinant DNA technology. The plasmids isolated from bacteria have been modified to easily carry out successful cloning experiments. These modified plasmids can be purchased from various suppliers as plasmid DNA vectors. The gene of the recombinant protein can be inserted in the so-called cloning region of the vector containing multiple cloning sites. There are several special unique sequence elements recognized and cleaved by restriction endonucleases in this region, as it is shown by the example of the pUC18/19 vectors in **Fig. 47**. The information about the DNA sequence of the vectors and the cloning regions can be obtained from the suppliers. Analysing the sequences is necessary for the success of the experiment.

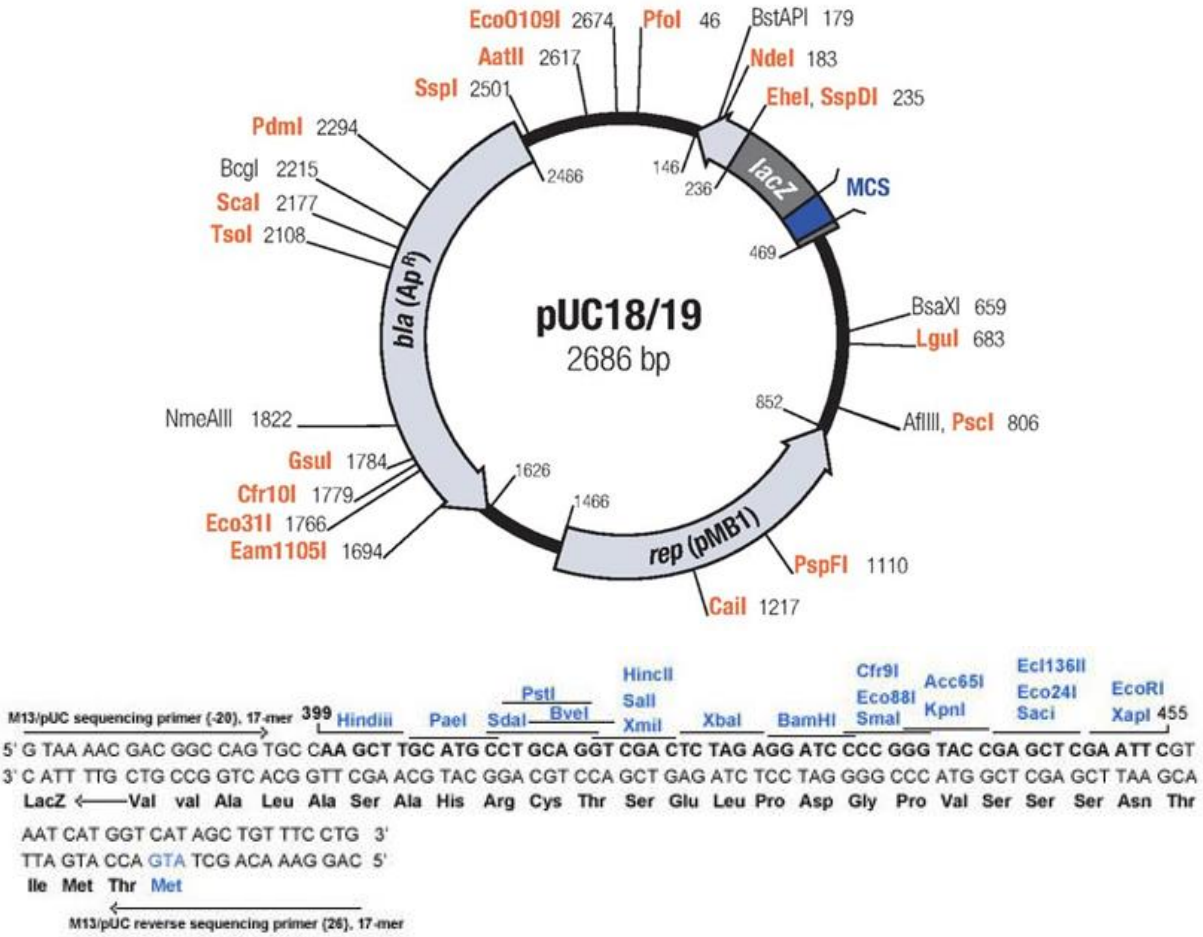


Figure 47. The schematic of the pUC18/19 vectors showing the various properties of these carriers for DNA cloning. This kind of representation is usually called a vector map. The region denoted by the MCS abbreviation is the multiple cloning site. This region is detailed in the bottom part of the figure showing the restriction enzyme recognition sequences within the MCS of pUC18 vector. pUC19 is similar to pUC18, but the MCS region is reversed. In the name of plasmids, such as also the pUC19, the "p" prefix denotes plasmid. Here the abbreviation UC stems for the University of California, where early work on the plasmid series had been conducted by its developers, Joachim Messing and co-workers.

Based on the available restriction sites in the cloning region, the selected nucleases can be used to cut plasmid DNA leaving e.g. sticky ends, at the cleavage

site. The termini cut by the same enzyme match specifically. Thus, the same selected restriction sites, built in by the primers into the gene encoding the target protein can also be cleaved. Then the gene can be inserted into the plasmid via the matching termini. These termini are then covalently linked by the ligase enzyme. This procedure is schematically depicted in **Fig. 48.**

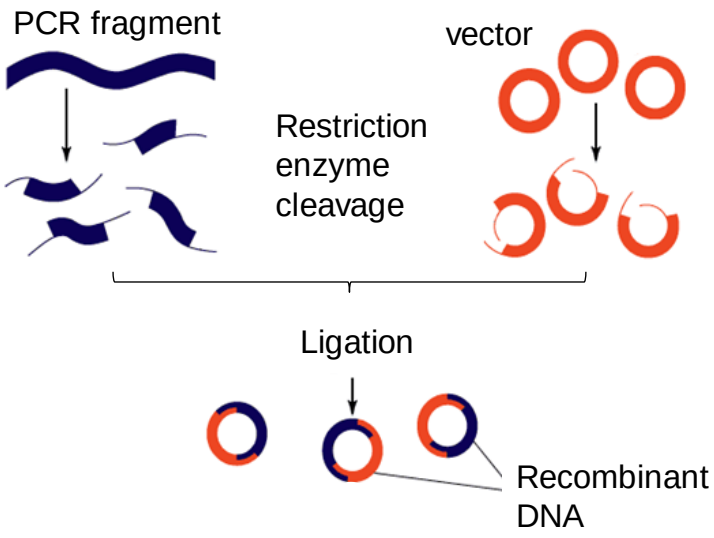


Figure 48. Schematic depiction of the construction of a recombinant DNA from the amplified PCR fragments and the plasmid DNA vector. Both types of DNA are cleaved with the same restriction enzyme and then joined into a single circular DNA by the help of the sticky ends and the ligase enzyme.

Multiple choices of the cloning vectors with various properties are available, so that the optimal strategy can be established for each specific cloning task. Nevertheless, the reaction depicted in **Fig. 48.** is not as simple.

The ligase enzymes are not very efficient, and also there is a possibility for multiple side products. The identical DNA molecules can be ligated to each other

forming dimers or oligomers; by the ligation of two different molecules linear heterodimers may also form; individual DNA molecules can be self-ligated to form circular DNA; the remaining fragments cleaved off from the PCR products may also interfere with the ligation if they are not removed prior to the experiment, etc. The only favourable outcome of the reaction is the formation of the circular DNA, containing one plasmid and one PCR fragment. For this reason, the ligation reaction has to be optimized as much as it is possible. To enhance the ligase catalysis specifically prepared ligase reaction mixtures, including secret additives are sold, which have to be mixed at certain volume ratio with the cleaved DNA fragments to be coupled. It is also important to mix the plasmid and the PCR fragment in appropriate ratio. For optimal reaction result, usually ~1:5 molar ratio of the plasmid and PCR fragment is advised. Therefore, the concentration of the cleaved DNA has to be determined. Nowadays, the easiest concentration determination of DNA is carried out by spectrophotometry. Based on the specific absorbance of double strand DNA at 260 nm (1 absorbance unit corresponds to a dsDNA solution of 50 μg/ml, or in other words, the mass extinction coefficient of dsDNA at 260 nm is 0.020 $(\mu g/ml)^{-1}$ cm^{-1}), the concentration can be determined from absorbance measurements. Furthermore, the spectrophotometry has the ability to verify sample purity using the A_{260} / A_{280} value (i.e. the ratio of the absorbances obtained at 260 and 280 nm, respectively). For pure DNA this value is considered to be between 1.8 and 2.0. The measured value below 1.8 suggests aromatic impurities, such as proteins of remaining phenol (form the protein extraction experiment). If the A_{260} / A_{280} ratio is higher than two, the DNA preparation usually contains RNA impurity. The absolute value of the recorded A_{260} should be in the proper range, so that it shall obey the Lambert-Beer law. If

this is not the case a misleading result is obtained. The specific property of the DNA samples, that only a small amount is available for experiments. Thus the common 1 cm long cuvettes can not be used in most of the cases. The new spectrophotometric instruments developed for DNA, RNA and protein concentration determination use only a 1 μL drop for the measurement. Such an instrument is shown in **Fig. 49**. Such a simple instrument measuring at two wavelengths requires careful sample purification. Only the concentration of pure samples can be determined accurately. More advanced NanoDrop instruments can record the full UV spectrum, providing further information on sample.

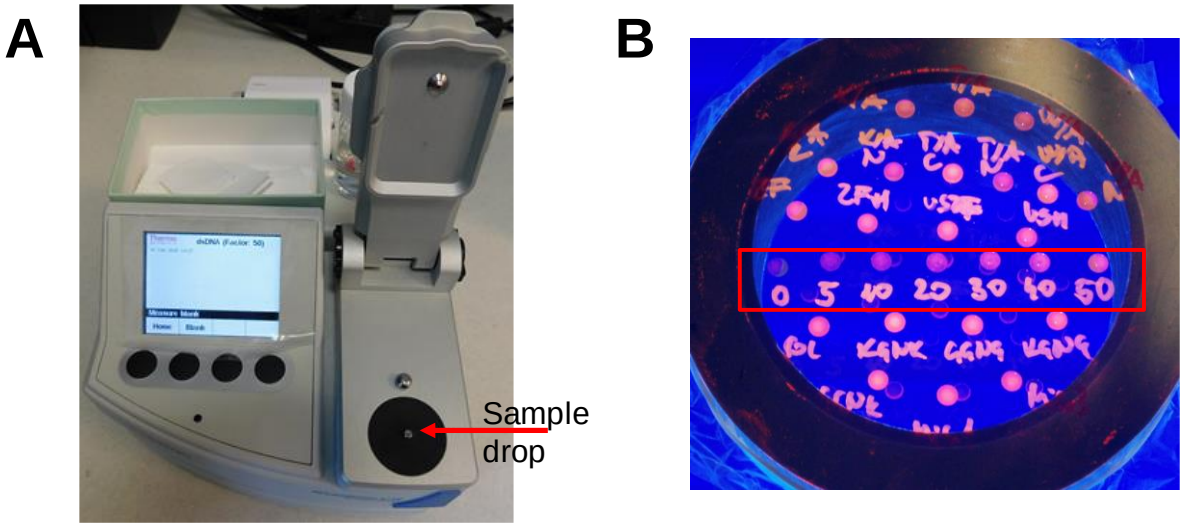


Figure 49. A) The NanoDrop instrument used in the laboratory of the author for measuring the absorbance of a drop of the DNA analyte at 260 and 280 nm. **B**) Concentration estimation by spot analysis of the DNA using ethidium bromide as a fluorescent agent. The red frame is used here to show the "calibration series" of the DNA spots for comparison with the fluorescence intensity of the spots of the analytes.

The spot analysis of DNA can also be used for estimation of the DNA concentration if a spectrophotometer is not available at the laboratory, or if for some reason it is not possible to use. In this method, spots are created of the same final total volume, containing the same amount of ethidium bromide. Then certain volume of the DNA is added and the final volume is adjusted by water. Illuminating the spots with UV light the fluorescence of the analyte spot can be compared to the intensity of the spot series created for calibration of the method.

To minimize the side reactions in the ligation experiment, it is also possible to cleave the PCR fragments and the plasmid with two different restriction endonucleases. In this way the two termini are different thus, the self ligation can be excluded, or will be negligible. The corresponding sticky ends created by the same enzyme can then only be ligated to each other, increasing the probability of the formation of the expected product. Such a cleavage experiment is shown in **Fig. 50**.

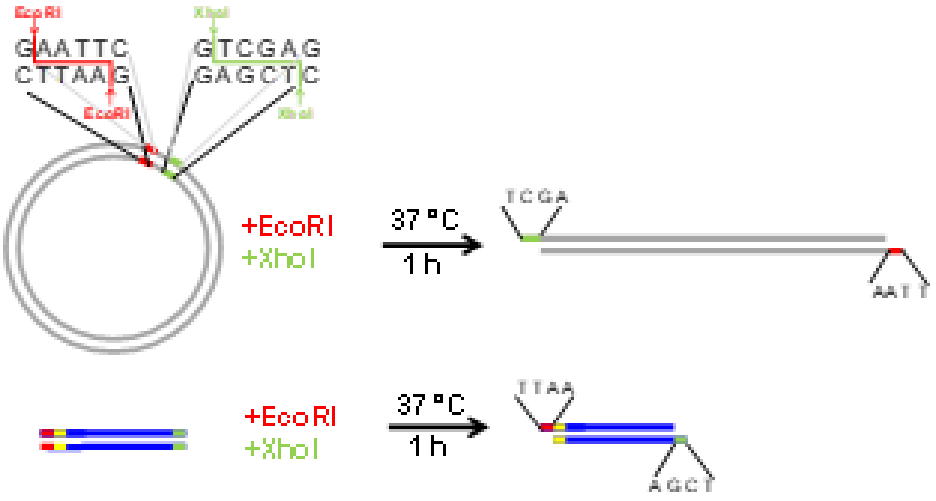


Figure 50. Cleavage of the plasmid (grey) and the PCR product (blue) by two restriction endonucleases: EcoRI and XhoI. (Taken from the PhD dissertation of Eszter Németh, written in the laboratory of the author of this e-book.)

After carrying out the ligation experiment the ligated DNA reaction mixture is added to an *E. coli* bacterial suspension. There are several bacterial strains in frequent use for the purpose of DNA cloning. These strains are already optimized for safe laboratory work, but the appropriate precautions have to be kept, during the laboratory work. In order to make bacteria susceptible for the plasmid DNA, they have to be treated in a specific way. The treated bacteria are called competent bacteria. Such *E. coli* bacteria will internalize the plasmid DNA, the process called transformation. Then the plasmid will be multiplied during the cell division: identical plasmid DNA molecules (clones) will be produced by the bacteria.

Next, the bacterial suspension will be spread to an LB agar-agar plate. LB is a rich medium for culturing bacteria. LB is often explained as the abbreviation of Luria-Bertani, the researchers working together for the time, when the first recipe of the LB medium was published. However, the original meaning of the abbreviation was: lysogeny broth. Incubating the plate overnight at 37 °bacterial colonies will grow on the plate. To avoid the growth of those bacteria, which were not transformed, i.e. did not internalize the plasmid, a specific antibiotic is mixed into the LB medium. Depending on the plasmid, various antibiotics are applied. The pUC18/19 plasmids carry a gene, which is responsible for the ampicillin resistance of the transformed bacteria. The principle of this is very simple. The transformed bacteria can express a protein, encoded by the plasmid, which can degrade the antibiotics. Ampicillin (**Fig. 51.**) and other β-lactam type antibiotics (such as the penicillin, amoxicillin, etc.) can be degraded by the β-lactamase enzyme. Thus, the bacteria expressing this protein will survive on the LB(amp+) plate (amp+ means that the LB medium used for the preparation of plates contains ampicillin).

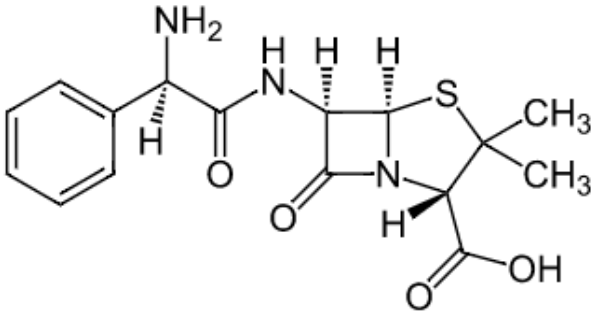


Figure 51. The structural formula of ampicillin: (2S,5R,6R)-6-[[(2R)-2-amino-2-phenylacetyl]amino]-3,3-dimethyl-7-oxo-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylic acid. Find the β-lactam part of the molecule.

The viable transformed bacteria form colonies on the plate, as it is shown in **Fig. 52**. It is important to notice that beside the ligation experiment, at least two control experiments were carried out here to make sure that the right conclusion can be drawn from the result.

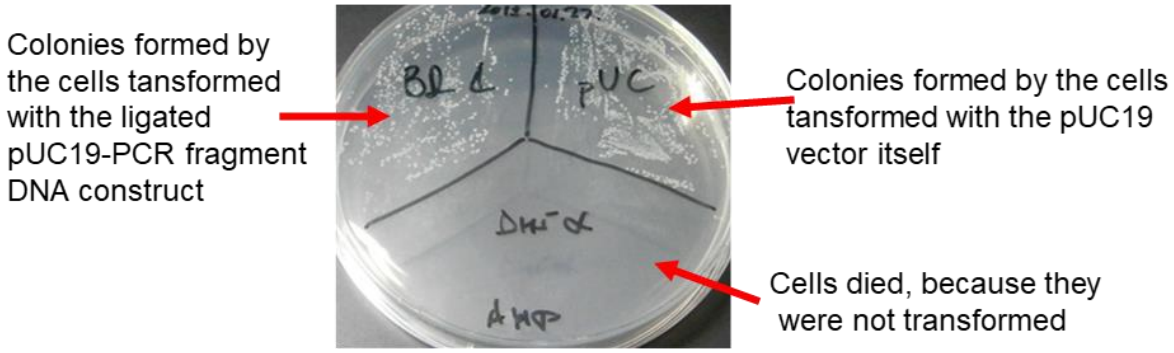


Figure 52. The LB(amp+) plate after an overnight incubation. Bacterial colonies were formed in those parts of the plate, where transformed cells were spread.

As the negative control, bacterial suspension without transforming them, was also spread on the plate. The fact that there are no colonies on this area of the plate demonstrates that the ampicillin was functional and the bacteria themselves were not resistant to the antibiotic.

The bacteria from the suspension mixed with the pUC19 plasmid itself were able to survive and form colonies. This part of the plate served therefore, as the positive control. By means of it, the correctness of the transformation procedure was approved.

Therefore, it is guaranteed that on that part of the plate, where the bacteria transformed with the target recombinant DNA were spread, only cells expressing the antibiotics could survive. However, there might also be bacterial colonies which only contain the pUC19 plasmid as the result of the self ligation of the vector. This has to be verified.

It is important to discuss here shortly one more property of the cloning plasmid vectors. It has already been mentioned that all these vectors possess a multiple cloning region for the flexibility of the selection of restriction sites in the cloning experiment. They also carry one or more genes encoding for proteins responsible for antibiotic resistance. These genes are called selection markers, since they allow for selection of bacteria, based on their resistance towards the antibiotics. The third property to introduce is the so-called origin of replication. It is a specific sequence in a genome at which replication is initiated. In principle there might be more different plasmids within one bacterial cell, e.g. the pUC19 plasmid with and without the inserted PCR fragment. However, the plasmids, which belong to the same incompatibility group can not exist and replicate in the same cell. The incompatibility group depends on the type of the origin of

replication. Plasmids with the same origin of replication are incompatible. Since the pUC19 plasmids with and without the inserted PCR fragment possess the same origin of replication, in practice, they can not exist and replicate in parallel within one bacterial cell. It also has to be mentioned that each bacterial colony on the plate is originated from a single bacterial cell. Thus, the above conclusion is valid for the bacterial colonies, as well. This is a property of the bacterial cells, by means of which we can distinguish between DNA fragments inserted into a DNA vector, independently of their size and sequence. As it was mentioned in Chapter 2, by means of this method a mixture of very similar DNA molecules can unequivocally be separated from each other. As an exercise, try to establish a strategy for such a separation experiment.

Various artificial plasmids carry even more advanced strategies to make the cloning and identification more simple and easy. They are of great importance in the course of the recombinant DNA technology. It is not surprising that Nobel Prize was awarded for the pioneering work on this field, as well (**Fig. 53**.).

Figure 53. The Nobel Prize in Chemistry in 1980 was awarded to Paul Berg "for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant-DNA" (Photo from the Nobel Foundation archive.)

Furthermore, not only plasmids, but other types of DNA molecules can also behave as vectors, i.e. carriers of a selected gene. The selection of these vectors depends on e.g. the size of the DNA fragment to be inserted, the organism used for cloning, the transformation procedure, and many more factors. Some examples are shown in **Table 7**.

Table 7. Examples of systems used for DNA cloning experiments:

Organism	**Vector**	**Method of delivery**
E. coli	plasmid	transformation, electroporation
	bacteriophage	phage injects the genetic material
	cosmid	entering the cells as a phage but and multiplying as a plasmid
	bacterial artificial chromosome	electroporation
Yeast	yeast artificial chromosome	electroporation
Eucaryotic cell	viruses	transfection
	artificial chromosome	transfection

After cultivation of the selected colonies, the plasmid DNA can be extracted by alkalic lysis of the bacteria, followed by a column purification and ethanol precipitation. The success of the insertion of the target gene into the plasmid needs to be checked by agarose gel electrophoresis and DNA sequencing. **Fig. 54.** e.g. shows the result of an AGE experiment, which was carried out on 12 purified

plasmid samples for the verification of the success of the ligation experiment. The upper part shows the bands of the plasmids. In every case two bands are visible. The size of the plasmid DNA with and without the insert differs in less than 10% of the size of the plasmid. Since the separation of these relatively large DNA molecules is ambiguous, the restriction endonucleases have been invoked in the experiment visualized in the bottom part of the figure. In this experiment the plasmid aliquots were cut by two restriction enzymes, which cleave the original plasmid into two linear DNA of equal size, one of which contains the multiple cloning site. As the consequence of this cleavage, a single band will appear on the gel at the half size of the plasmid DNA.

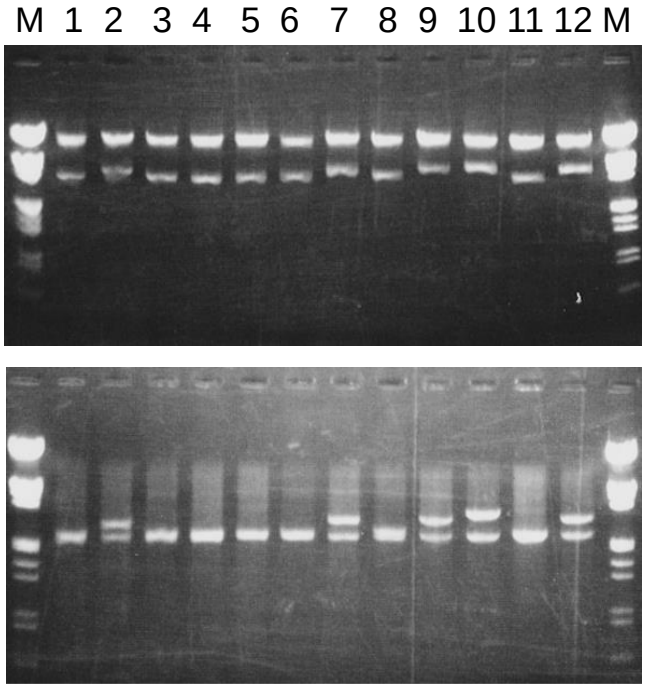


Figure 54. The photo of the agarose gel, as the result of an AGE experiment: 12 purified plasmid samples were run on the gel to verify the success of the ligation experiment. M denotes the DNA marker.

However, the successful insertion of the PCR fragment will make difference between the sizes of the two restriction fragments. Thus, two bands will appear on the gel unambiguously indicating the insertion of the DNA fragment. This is the case in lanes 2, 7, 9 and 12. It is interesting to note that two bands are also observed in lane 10, but the upper band indicated somewhat larger DNA fragment than in the other successful experiments. One can speculate, that the insert might be the dimer of the PCR fragment (if only one restriction endonuclease was used in the cloning procedure).

Based on this result, the precise DNA sequence with the emphasis on the multiple cloning site of the plasmids run in lanes 2, 7, 9 and 12, have to be identified. This is carried out by DNA sequencing procedure.

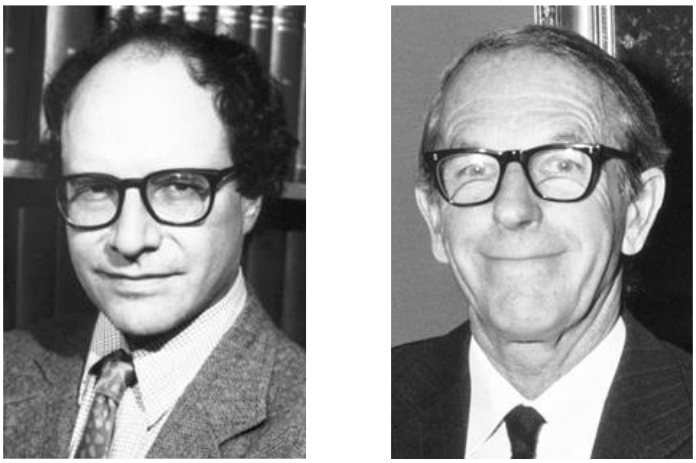


Figure 55. The Nobel Prize in Chemistry was awarded in 1980 jointly to Walter Gilbert and Frederick Sanger (from left to right) "for their contributions concerning the determination of base sequences in nucleic acids." (Photo from the Nobel Foundation archive.)

Since the DNA is a quite uniform molecule, taking into account that all the 2'-deoxynucleotides have very similar chemical properties, the sequencing seems to be an extremely difficult task. Nevertheless, the invention of PCR initiated the development of a simple procedure for sequencing, which in fact also deserved Nobel Prize in Chemistry, awarded jointly to Walter Gilbert and Frederick Sanger (**Fig. 55.**).

As the conclusion of this chapter and because of its importance, the DNA cloning procedure in its broad sense is summarized in **Fig. 56.**

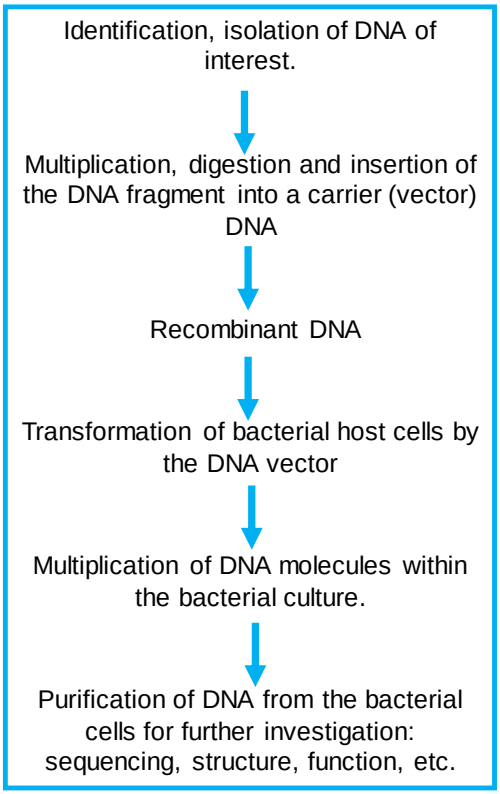


Figure 56. The summary of the DNA cloning procedure, starting from the selection of the target gene and terminating with the unambiguous identification of the cloned DNA fragment.

Monitoring questions

- Explain the meaning of DNA cloning?

- What are restriction endonucleases? What are they used for in DNA cloning?

- What is the restriction/methylation system?

- What is the restriction site? What kind of restriction sites are recognized by Type II restriction endonucleases?

- How is the DNA cleaved by the restriction endonucleases?

- What are palindromic sequences in terms of the double strand DNA?

- What are the guidelines of the selection for restriction site(s) in a DNA cloning experiment?

- What is the advantage of using two different restriction enzymes for DNA cloning? Imagine, what are disadvantages?

- How the sticky ends are formed during the restriction cleavage of the DNA?

- What is the average length of the DNA fragments resulting from the XhoI restriction endonuclease treatment of a long DNA?

- Which factors influence the efficiency of the cleavage of the PCR fragments by restriction endonucleases?

- What is the definition of the unit of the restriction endonuclease?

- Which enzyme can couple two DNA fragments covalently?

- How can be the concentration of the DNA solution determined? Which property of the DNA makes it possible that a general specific absorbance can be applied for each DNA solution?

- Explain the meaning of the cloning vector. What kind of cloning vectors are mentioned above?

- Which are the three most important characteristics of the cloning vectors.

- What is the multiple cloning site used for?

- What is the meaning of the selection marker?

- What is the importance of origin of replication?

- What kind of control experiment should be performed during the transformation of the bacterial cells?

- How can be DNA molecules separated by the help of bacterial cells?

8. Transcription and translation

Students who study this chapter will acquire the following specified learning outcomes:

Knowledge

The students understand the concepts of transcription and translation in terms of biochemistry.

The students know the meaning and role of the operon.

The students are aware of various types of agents inducing the transcription.

The students list various types of RNAs and explain their biological function.

The students understand the role of the aminoacyl-tRNA synthetases in deciphering the genetic code.

The students know the principles of the protein synthesis in ribosome.

Skills

The students explain the regulatory mechanism of the transcription through the operon.

The students make difference between the cloning and expression DNA vectors.

The students understand the functioning of T7 promoter in E. coli BL21(DE3) cells.

The students know the structure of RNA and analyse the consequences of the structural differences between DNA and RNA.

The students design an operon including more genes.

Attitude

The students pay attention to the importance of correct design of the oligonucleotide primers avoiding frame shifted genes.

The students make effort to widen their knowledge on protein synthesis in ribosomes analysing crystal structures recently published in the scientific literature (Science, Nature)

Responsibility and autonomy

The students translate the DNA sequence independently into protein sequence and vice versa.

The students discuss their design of the operon with their colleagues.

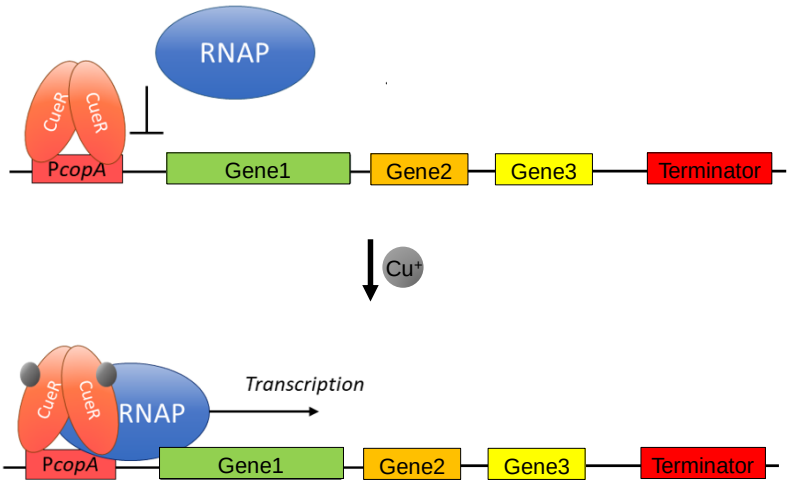
The recombinant DNA, as synthesized in the previous chapter is ready for protein expression. However, instead of the cloning vectors, expression vectors are used for this purpose. Nowadays, these two properties can be combined together in the artificial plasmids. If not, the gene has to be recloned from the cloning vector into an expression vector. The latter vectors possess the same major characteristics, as it was listed for the cloning vectors: (i) there is a multicloning site in the vector with selected unique restriction sites; (ii) there is one or more gene encoding for enzymes providing the antibiotics resistance – selection markers, and (iii) there is a specific origin of replication in each vector.

The main difference between the two types of the vectors is that the multiple cloning site is inserted between regions called promoter and terminator sequences in the expression vectors. The DNA sequence including the promoter, the terminator sequence and the region between these two is called operon. This region of the DNA is responsible for the regulation of the first step of the information transfer from DNA to protein, which is the synthesis of the mRNA molecules from a DNA template, called transcription. In this process the RNA polymerase is copying the DNA strand encoding for the target protein into an RNA molecule. The promoter sequence serves as the RNA polymerase binding site, while the terminator sequence is a signal for the enzyme to stop the transcription process.

Since the proteins encoded by the genome of an organism are not necessarily present in the cells during its whole lifecycle, this process shall be regulated. This usually occurs through the so-called operator region, to which various molecules, e.g. proteins can bind to either prevent or stimulate the

transcription process. Depending on the operons, various substances (small non-proteinic molecules, ions, metabolites) are able to bind to e.g. the repressor proteins activating the transcription.

A



B

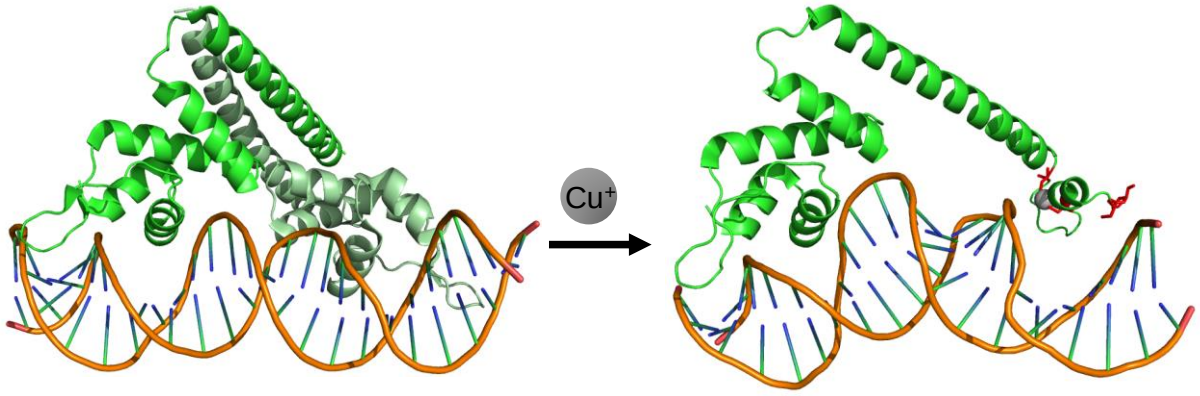


Figure 57. A) The schematic description of an operon, responsible for copper homeostasis in *E. coli*. P*copA* is the so-called *copA* promoter region. RNAP is the RNA polymerase. B) PyMol image of the apo and metal ion bound CueR protein in complex with DNA. The distortion of the DNA molecule is clearly visible upon metal ion binding. The figure was constructed based on the crystal structure coordinates with PDB Id: 4WLS and 4WLW. The protein molecules are depicted in green, while the DNA is in orange.

As an example, the operon including the genes of proteins, which are responsible for copper homeostasis of the bacterial cells is shown in **Fig. 57**. In this system, the CueR copper regulatory protein acts as a repressor. Binding to the operator, which is in this case overlapped with the promoter region of the DNA, it prevents the binding of the RNA polymerase (RNAP). Whenever unwanted copper excess appears in the cell, the Cu^{+} ions bind to CueR, causing a small conformational change, which results in DNA distortion, and makes the RNA polymerase binding possible. Thus, the enzyme can start the transcription process.

One or more genes can be included in a single operon. These genes will be transcribed in parallel. Thus, the proteins will also be synthesized in parallel. As an example, the simultaneous expression of the enzymes of the above mentioned restriction / methylation system can also be achieved through such a regulatory system. This strategy is also often applied by the cells, when expressing toxic proteins, which are synthesized in parallel with their immunity proteins.

The lac operon (lactose operon) was the first operon, the regulatory mechanism of which was described in detail. This operon is required for the transport and metabolism of lactose in *E. coli*. When the nutrient contains glucose, bacteria activate the glucose-metabolizing enzymes. There is no other active metabolic pathway under such conditions. Providing lactose instead of glucose, the lactose-metabolizing enzymes appear in the cell – so that lactose can be used as nutrient. The explanation of this adaptation to the new conditions can be provided by the lac operon model. β-galactosidase enzyme is expressed for digestion of lactose when glucose is not available as carbon source.

The lac operon thus, became the foremost example of prokaryotic gene regulation. Nobel Prize has been awarded to François Jacob and Jacques Monod for description of the lac operon (**Fig. 58.**).

Figure 58. The Nobel Prize in Physiology or Medicine 1965 was awarded to François Jacob and Jacques Monod (from left ot right) "for their discoveries concerning genetic control of enzyme and virus synthesis" (Photo from the Nobel Foundation archive.)

The lac operon is under double regulation. The amounts of glucose and lactose determine the extent of the transcription. Accordingly, four states of the operon can be distinguished as shown in **Table 8**. In the presence of glucose, both the lactose transport into the cells and the metabolism of the lactose is inhibited. The lac operon is only switched on, when there is a lack of glucose in the nutrient in parallel with the presence of lactose. The lac repressor, a protein bound to DNA sequence, containing the lac operon is released from the DNA when allolactose is bound to the repressor. This event initiates the transcription of the genes responsible for lactose metabolism.

Table 8. The four states of the lac operon.

Nutrient	**Transcription**
+glucose +lactose	OFF
+glucose –lactose	OFF
–glucose –lactose	OFF
–glucose +lactose	ON

Similarly, initiation of transcription occurs when e.g. isopropyl β-D-1-thiogalactopyranoside (IPTG) was added. IPTG is a molecular mimic of allolactose, a lactose metabolite that triggers transcription of the lac operon. However, the sulfur atom in IPTG is able to covalently bind the repressor, so that it prevents the cell from degrading the inducing agent. In this way, the IPTG concentration remains constant, and the transcription is continuously switched on. This leads to overproduction of the RNA and as the consequence, to the overexpression of the protein. This advantageous property of IPTG made it a commonly used inducing agent in protein expression experiments. With the expression vectors using the lac operon for transcription, this process is regulated by IPTG.

Similarly, *E.coli* BL21(DE3) bacterial strain is optimized for protein expression. These cells contain the gene of the bacteriophage T7 RNA polymerase on their chromosomal DNA, being under the control of the lac promoter.

Therefore, the transcription of the RNA and expression of T7 RNA polymerase can be achieved by adding IPTG to LB medium. The expressed T7 RNA polymerase initiates the transcription on any expression vector that contains the T7 promoter sequence (see **Fig. 59.**), finally resulting in the expression of the gene(s) under the control of this promoter. The bacteriophage T7 RNA polymerase is a popular enzyme for transcription of a plasmid DNA in *E. coli* BL21(DE3) cells.

Figure 59. PyMol image of the bacteriophage T7 RNA polymerase initiating the transcription of the DNA molecule into RNA. The protein is green, the DNA is orange and the growing RNA strand is pink. The figure was constructed based on the crystal structure coordinates downloaded from RCSB Protein Databank. PDB Id: 1MSW.

Several advantages of bacteriophage T7 RNA polymerase can be listed, such as the very high activity (it synthesizes RNA several times faster than the *E. coli* RNA polymerase), it terminates transcription less frequently, it is highly selective for initiation at its own promoter sequences, and it is resistant to antibiotics such as rifampicin, inhibiting the *E. coli* RNA polymerase. For these reasons, many expression vectors use the T7 promoter to control the protein production through the transcription.

Transcription thus, produces RNA molecules. These ribonucleic acids are somewhat different from DNA. They consist of ribonucleotide monomeric units possessing ribose instead of 2'-deoxyribose. I.e., a 2'-hydroxy group is also present in the ribonucleotide molecule. Being either a nucleophile, or a metal ion binding site, this group makes the RNA more sensitive to hydrolysis. The four types of nucleotides in RNA are abbreviated by A, U, G and C. U is uridine, which replaces the T, thymidine found in DNA. Uridine and thymidine differ in a single methyl substituent.

Various forms of the RNA molecules appear in the cells, with various functions, such as mRNAs (m = messenger), tRNAs (t = transfer), rRNAs (r = ribosomal), snRNAs (sn = small nuclear), snoRNAS (sno = small nucleolar), siRNAs (si = small interacting), miRNAs (mi = micro), etc. RNAs are single strand molecules in contrast to the double helix of DNA. Nevertheless, RNA molecules can form secondary structures by intramolecular interactions. **Fig. 60.** demonstrates that even small nuclear RNA molecules form secondary structures.

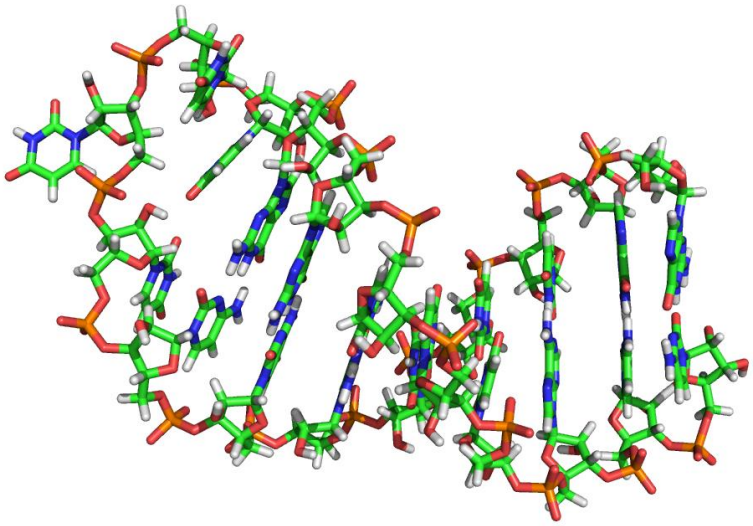


Figure 60. PyMol image of of U2 snRNA stem I from *S. cerevisiae* as determined by NMR. The figure was constructed based on the coordinates from RCSB Protein Databank. PDB Id: 2O33.

It shall be mentioned here that the coding region of the DNA in eukaryotes is not a continuous sequence in contrast to the prokaryotes. The gene in a eukaryotic cell consists of coding exons and non coding introns. The primary transcript contains the copy of the whole sequence between the promoter and terminator regions. Then, by the process called splicing the introns are cut out from this primary transcript resulting in the messenger RNA.

This mRNA serves as the template for the protein synthesis in the ribosomes. The process in which the protein molecules are synthesized based on the mRNA code, is called translation. This process takes place in the ribosome, which is a sophisticated complex of several proteins and ribonucleic acids, collaborating with each other. The translation is based on the genetic code. One amino acid is encoded by a codon, which consists of three subsequent nucleotides in the RNA. Since there are altogether $4^3 = 64$ possible nucleotide triplets for 22

amino acids, some of the amino acids may even have multiple codes. The codes are collected in **Table 9**.

Table 9. The RNA triplets encoding for amino acids.

5' end					**3' end**
	U	**C**	**A**	**G**	
U	UUU = Phe UUC = Phe UUA = Leu UUG = Leu	UCU = Ser Ser Ser Ser	UAU = Tyr Tyr End End	UGU = Cys Cys End Trp	**U** **C** **A** **G**
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	**U** **C** **A** **G**
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	**U** **C** **A** **G**
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	**U** **C** **A** **G**

In fact, the key enzymes in deciphering the genetic code are the aminoacyl-tRNA synthetases. These are highly specific enzymes, which couple the appropriate tRNAs with their cognate amino acids. **Fig. 61.** shows the glutamyl-tRNA synthetase complexed with tRNA(Glu).

The appropriate amino acid is coupled to the 3' end of the tRNA by the enzyme, and the corresponding anticodon loop is highlighted by blue background and sticks. The anticodon has the following sequence in **Fig. 61**: 5'-CUC-3'. Its complementary sequence on the mRNA is 5'-GAG-3', coding for Glu according to the **Table 9**.

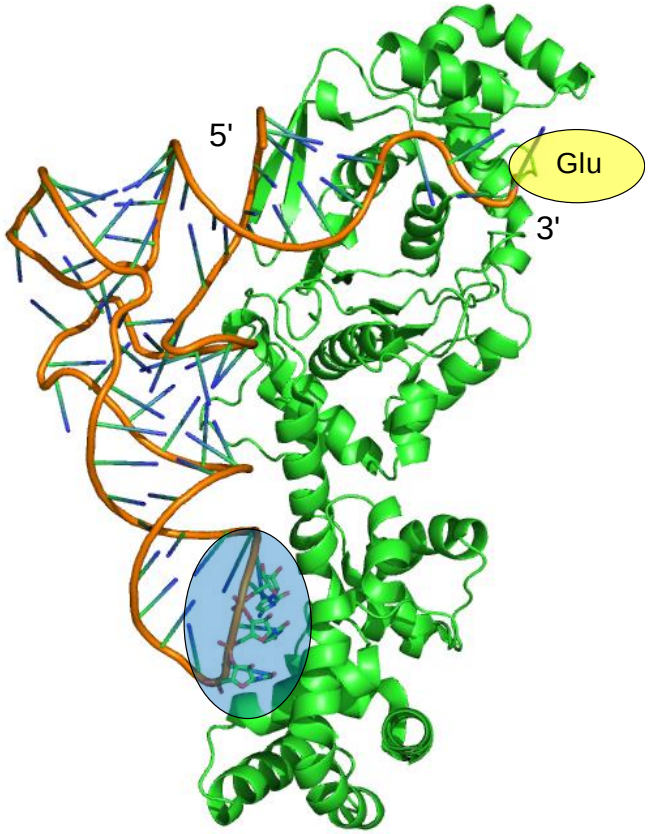


Figure 61. PyMol image of glutamyl-tRNA synthetase complexed with tRNA(Glu). The figure was constructed based on the crystal structure coordinates from RCSB Protein Databank. PDB Id: 1G59. The yellow ellipse symbolizes the attached amino acid to the 3' end of the RNA, while the anticodon loop is highlighted by blue background and sticks.

The mRNA is recognized by the ribosome through the ribosome binding site (RBS). In prokaryotes the RBS, also called the Shine-Dalgarno (SD) sequence is a consensus 5'-AGGAGG-3' sequence. Downstream, i.e. towards the 3' terminus of the RBS the 5'-AUG-3' start codon is located. Then, the bound mRNA serves to direct the amino acid loaded tRNAs to the site of the reaction. The mRNA and tRNA, both complexed with each other and with the ribosome provide

the framework for protein synthesis. The reacting groups of the growing protein chain and the incoming amino acid complexed with tRNA approach each other for the peptide bond formation to occur. This process is accompanied by the multiple conformational changes of the constituents of the ribosome.

A ribosome is a ribonucleoprotein consisting of RNAs and proteins. Each ribosome is divided into two subunits collaborating with each other: (i) a smaller subunit which binds the mRNA through base pairing with the ribosomal RNA, and (ii) a larger subunit which binds to the tRNA, being the site of the peptide bond forming reaction. *E. coli* bacteria have 70S ribosomes, consisting of the small (30S) and the large (50S) subunits. (S is the unit of measurement of the rate of sedimentation during centrifugation.) There are three tRNA binding sites in the ribosome: A, P and E. The A-site binds an incoming aminoacyl-tRNA, the anticodon of which matches the codon of the mRNA. Only if this tRNA is properly matched, it will be used for protein synthesis. The ribosome catalyzes the peptide bond formation with the peptidyl-tRNA (the tRNA bound to the growing polypeptide chain) bound in the P-site. The peptide chain is transferred in this way to the incoming aminoacyl-tRNA. This is accompanied by a large conformational change in the ribosome. As the consequence, the free tRNA (the one, which released the peptide chain) is moved to the E site, while the tRNA bound to the peptide is moved to the P-site. Then, the A-site can bind the next incoming aminoacyl-tRNA and the procedure repeats until a stop (end) codon is met in the mRNA sequence. These steps are modelled in **Fig. 62**.

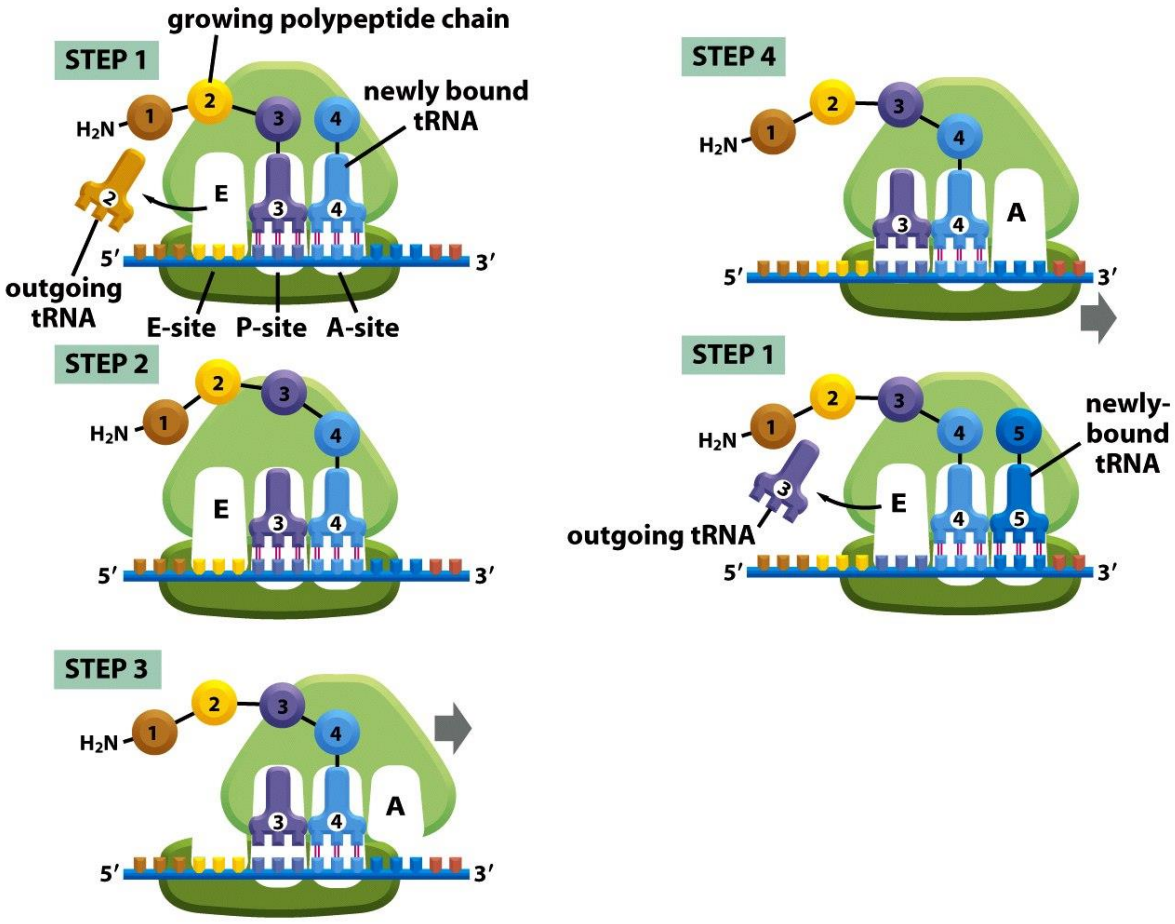


Figure 62. The schematic representation of the protein synthesis in the ribosome. The figure is taken from the The molecular biology of the cell, Garland Publishing Inc, New York, London, 1989.

In the ribosome the protein sequence is read from the start codon. Then every bse triplet is read as the appropriate amino acid, until the stop codon is reached. In an operon containing more than one gene, these are usually frame shifted thus, all of them require their own RBS sequences and start codons. **Fig. 63**. demonstrates the importance of the application of the correct reading frame.

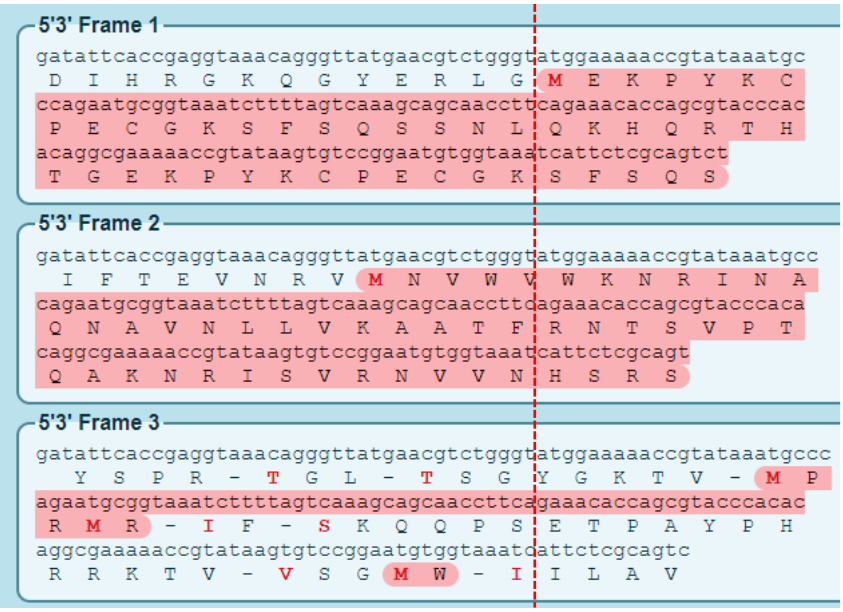


Figure 63. The translation of a DNA sequence in the three different 5' → 3' reading frames. The dotted red line is positioned at the first nucleotide of the 5'-ATG-3' start codon (5'-AUG-3' in mRNA). The translation was carried out using the Translate tool at the ExPASy Bioinformatics Resource Portal (https://web.expasy.org/translate/).

It can be concluded from **Fig. 63.** that by shifting the reading frame either a protein of different sequence is synthesized, or more probably stop codons will appear soon in the sequence resulting in a short peptide fragment, which is usually degraded by the cells. In theory the complementary strand of the DNA can also encode for a protein thus, it can also be translated, as shown in **Fig. 64.**

None of the reading frames on the complementary DNA strand could be translated into a long continuous protein sequence. These results demonstrate the importance of the reading frame adjustment or shift. Care has to be taken during the selection of the restriction sites to keep the proper reading frames. Mutations as small deletions or insertions can also cause the shift of the reading frame. In

living organism such a DNA modification may result in an inherited or cancerous disease, while in the laboratory, to a wrong experiment.

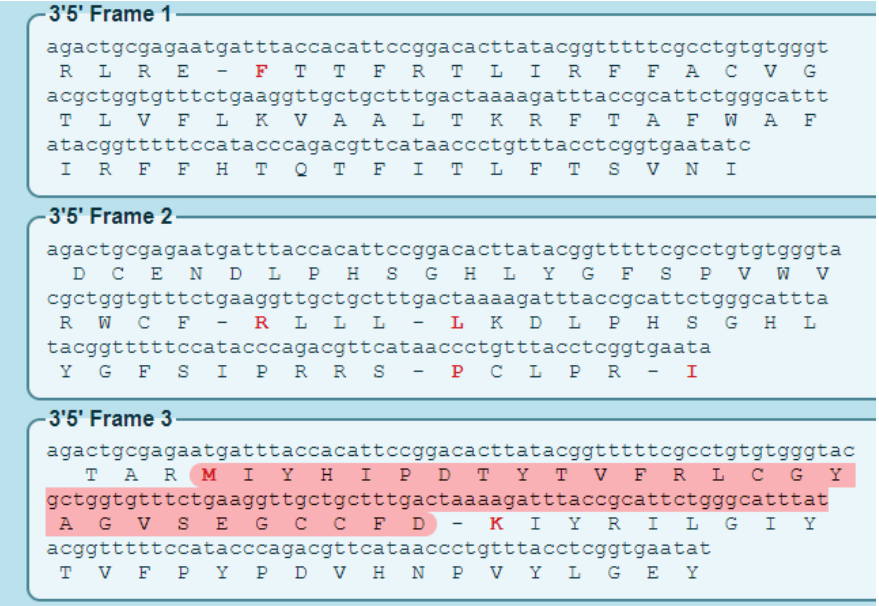


Figure 64. The translation of a DNA sequence in the three different 3' → 5', i.e. complementary strand reading frames. The translation was carried out using the Translate tool at the ExPASy Bioinformatics Resource Portal (https://web.expasy.org/translate/).

The proteins usually fold into their three dimensional structures immediately after the synthesis, but is some cases further processing and aid is needed to obtain the functional structure. When talking about the structure of the proteins four main structural levels are distinguished:

- The ***primary structure*** is the amino acid sequence of the protein, which is written from the N-terminal containing free α-amino group towards the C-terminal residue containing the free α-carboxylic group from left to right. The proteins expressed consist of amino acids bound together through peptide bonds. They

may contain hundreds of amino acids therefore, their sequence is usually written by using one letter codes, as already applied in **Fig. 63.** and **Fig. 64.** For identification of these characters, the codes are listed in **Table 10**.

Table 10. The corresponding one and three letter codes of amino acids.

A – Ala	E – Glu	I – Ile	N – Asn	S – Ser	Y – Tyr
B – Asx	F –Phe	K – Lys	P – Pro	T – Thr	Z – Glx
C – Cys	G – Gly	L – Leu	Q – Gln	V – Val	
D – Asp	H – His	M – Met	R – Arg	W – Trp	

- The ***secondary structure*** is formed by hydrogen bonding between the peptide nitrogens and oxygens of peptide bonds at various distances. Based on this, various helices, strands and turns are usually distinguished. Motifs and domains are also mentioned as supersecondary structures.
- The ***tertiary structure*** corresponds to the three dimensional structure of the protein chain.
- The ***quaternary structure*** is related to complex proteins consisting of more than one polypeptide chains. It is thus the three-dimensional structure of multimeric proteins.

The visualization of three dimensional structures of proteins is extremely useful in the understanding of their function and interactions. Drug molecules as enzyme inhibitors, receptor binders, etc are designed using such 3D structures. The structure of macromolecules, such as proteins can be determined at atomic details e.g. by X-Ray crystallography or NMR-spectroscopy. These structures can

be understood via visualizing them in three dimensions by using the Cartesian atomic coordinates, deposited in most of the cases into the RSCB Protein Data Bank (https://www.rcsb.org/) and freely available for downloads. Numerous softwares are able to visualize the molecules based on the list of coordinates of the atoms in the molecule. The figures in this e-book were created by the PyMOL 1.3 (The PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC.) free software for academic use. It has a graphical interface and command line as well. Representation of structures is available in different modes (lines, sticks, cartoon, surface etc.). The program can also perform the alignment of selected structures based on structural similarity. The root mean square deviation (RMSD) characterizing the superposed structures ca also be computed.

Monitoring questions

- Describe the main difference between the cloning and expression vectors!

- What is the operon?

- What is the role of the operator region of the DNA?

- What is the meaning of the word "transcription" in biochemistry?

- What is the main role of the lac operon? Under which condition is the lac operon "switched on"?

- Who has received Nobel Prize for the description of the lac operon? Learn more about the work of these scientists!

- What is isopropyl β-D-1-thiogalactopyranoside (IPTG) used for in recombinant DNA technology?

- Explain, the mechanism of transcription regulation of the T7 promoter containing vectors by IPTG in *E. coli* BL21(DE3) bacterial cells!

- What are the main differences between the structures of DNA and RNA molecules? Find the connection between the differences in the structure and differences in the function.

- List few types of RNA molecules and explain

- What is the meaning of the word "translation" in biochemistry?

- Describe the principles of the genetic code.

- Who was awarded Nobel prize for the breaking of the genetic code?

- Which are the two most important steps of deciphering the genetic code?

- Describe briefly the process of the protein synthesis in the ribosome.

- What is the consequence of a reading frame shift?

- What are the various levels in the hierarchic description of the protein structure?

9. Protein identification and purification

Students who study this chapter will acquire the following specified learning outcomes:

Knowledge

The students understand the concepts of primer design.

They are aware of the differences between DNA and protein gel electrophoresis methods.

The students understand the concept of the fusion tags and their use in protein purification.

The students understand the principles of the chromatographic methods of the protein purification.

The students know the hierarchy of the protein structure levels.

Skills

The students compare chromatographic protein purification strategies and select the appropriate method for their experiment.

The students analyse the results of the SDS-PAGE experiment in context of the success of the protein expression, protein amount and purity.

The students optimize the conditions for their protein expression and purification experiments

The students realize the context of the biological tool in chemistry as a whole unit.

Attitude

The students pay attention to the importance of correct design of the oligonucleotide primers in context of the choice of the protein purification strategy.

The students try to think about their experiment as a whole, to design the individual steps correctly.

The students pay attention to the design of the comparative experiments.

The students write and follow their protocols precisely.

The students take effort to read and comply with the requirements of the laboratory experiments in terms of safe and sterile work.

Responsibility and autonomy

The students build up a strategy for protein purification independently.

The students discuss their results with each other, with the emphasis on their colleagues from various research fields.

The proteins expressed in the cells are obtained as a mixture of many different proteins. Thus, it is essential to identify the target protein in this mixture to decide about the success of the experiment. Similarly to DNA identification, the gel electrophoresis seems to be the easiest method for this purpose. However, while the different DNA molecules have a uniform negative charge density, the charge of the proteins may vary according to the protonation state of the appropriate side-chains of the amino acid residues in the protein. The acidic side-chains, such as the carboxylic groups of aspartyl and glutamyl residues carry negative charge at neutral pH. The basic side-chains, such as the amino group of the lysine, and guanidine group of the arginine are protonated around pH ~ 7 thus, they possess positive charges. The side-chains of histidyl residues in proteins (de)protonate around pH ~ 6.5 thus, it is difficult to decide about their contribution to the overall charge. The overall charge of protein is determined by the sum of the charged groups. Since the protonation state is varying by pH it is possible to adjust the pH to obtain a net zero charge (i.e. the number of negatively and positively charged groups is equal). This pH is the isoelectric point (pI) of the proteins.

The gel electrophoresis of proteins can be performed taking into account their pI, i.e. by the so-called isoelectric focusing method. Nevertheless, it is also possible to carry out similar simple electrophoretic experiment as done for DNA. In this case the polyacrylamide gel electrophoresis (PAGE) is performed in the presence of sodium dodecyl sulfate (SDS). This SDS-PAGE experiment is an easy and quick way to detect the protein content and purity of the sample. The gel is prepared with the radical polymerization of acrylamide, with the addition of

N,N'-methylenebisacrylamide to crosslink the polymer. The reaction is initiated by ammonium-persulfate and controlled by N,N,N',N'-tetramethylethylene-diamine. The electrophoresis is carried out in a vertical arrangement (**Fig. 65.**). The gel consists of two layers: a short stacking gel (e.g. 6% acrylamide, pH = 6.8) and a long resolving gel (e.g. 12.5% acrylamide, pH = 8.8).

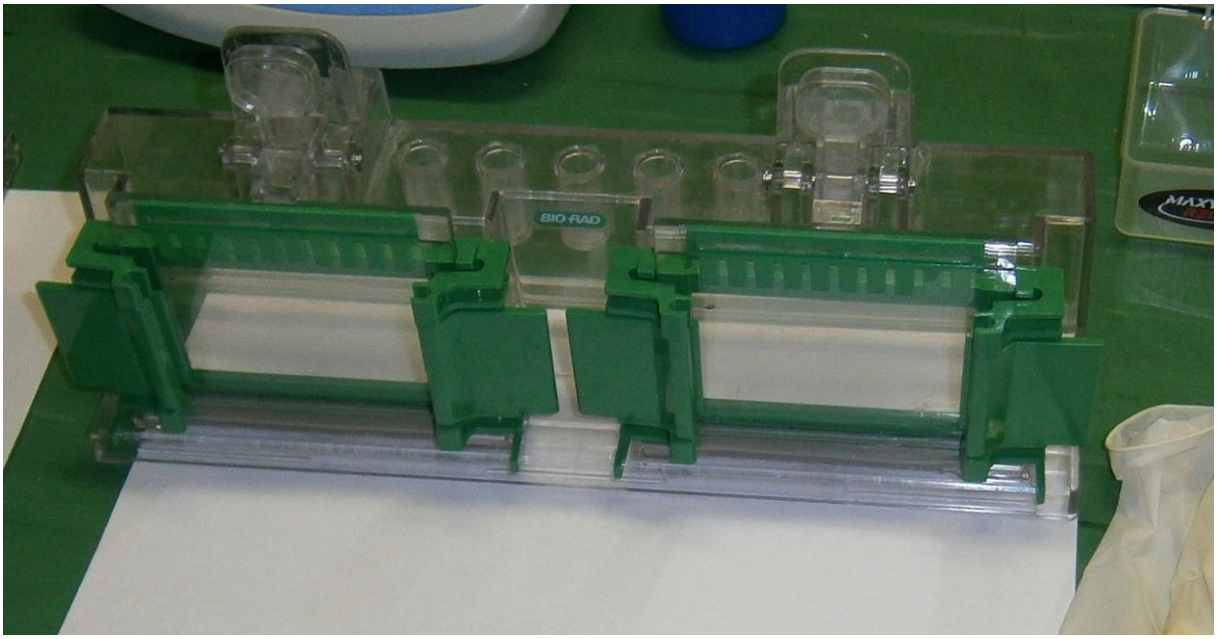


Figure 65. Casting the polyacrylamide gel in the laboratory of the author of this e-book, showing the vertical arrangement of the gel. It is important to note that gloves must be weared during the work with polyacrylamide – the safety precautions have to be studied carefully.

After the protein expression, the proteins are obtained from the cells by disrupting them usually by sonication. Such a sonicator is shown in **Fig. 66**. By means of this instrument, the bacterial cells can be disrupted in small volume Eppendorf tubes. Nevertheless, the treatment by ultrasound heats up the sample,

so the procedure has to be carried out on ice and without extensive formation of air bubbles to prevent the denaturation of the proteins.

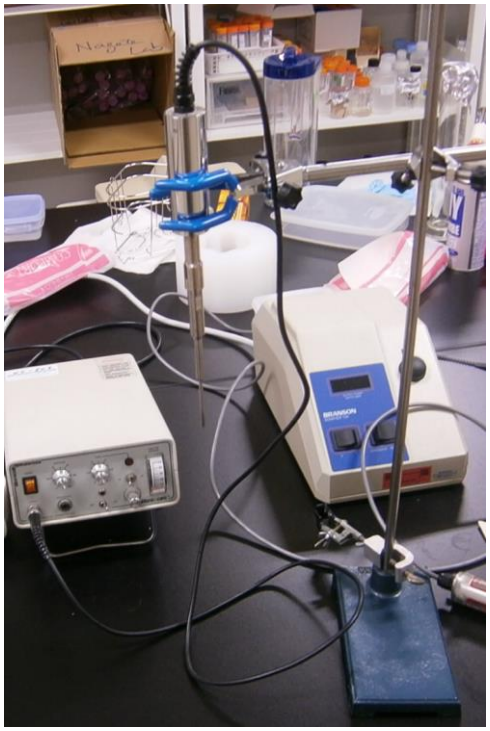


Figure 66. An example of a sonicator instrument used in the laboratory of the Japanese collaborator of the author of this e-book.

The initial samples are aliquots of the Total, Insoluble and Soluble fractions (see below in **Fig. 67.**). They are prepared for the electrophoresis by the adding a buffer (pH = 8.5) containing SDS and mercaptoethanol. SDS denatures the protein destroying the secondary interactions, while mercaptoethanol is a reducing agent for disulfide bridges. During the incubation at high temperature (~ 95 °C) for few minutes, the proteins are completely denatured. SDS binds to the hydrophobic regions, and the protein chains obtain negative charge, the amount of which is proportional to the length of the protein chain.

During the electrophoresis the proteins are first highly concentrated into narrow bands in stacking gel. Then, they migrate according to their molecular size, as their negative charge density due to the bound SDS molecules is uniform. The smaller proteins migrate faster, while the large proteins find their route through the labyrinths of the gel more slowly. Similarly to DNA, a loading dye helps the loading and monitoring the electrophoresis of the protein samples.

The protein bands can be visualized in the gel by staining with e.g. Comassie Brilliant Blue, an anionic triphenylmethane dye that nonspecifically binds proteins. Compared to molecular weight standards the size of the protein (length of sequence) and even its concentration can be estimated.

The result of such an SDS-PAGE experiment is depicted in **Fig. 67**. Here the success of the protein expression was followed.

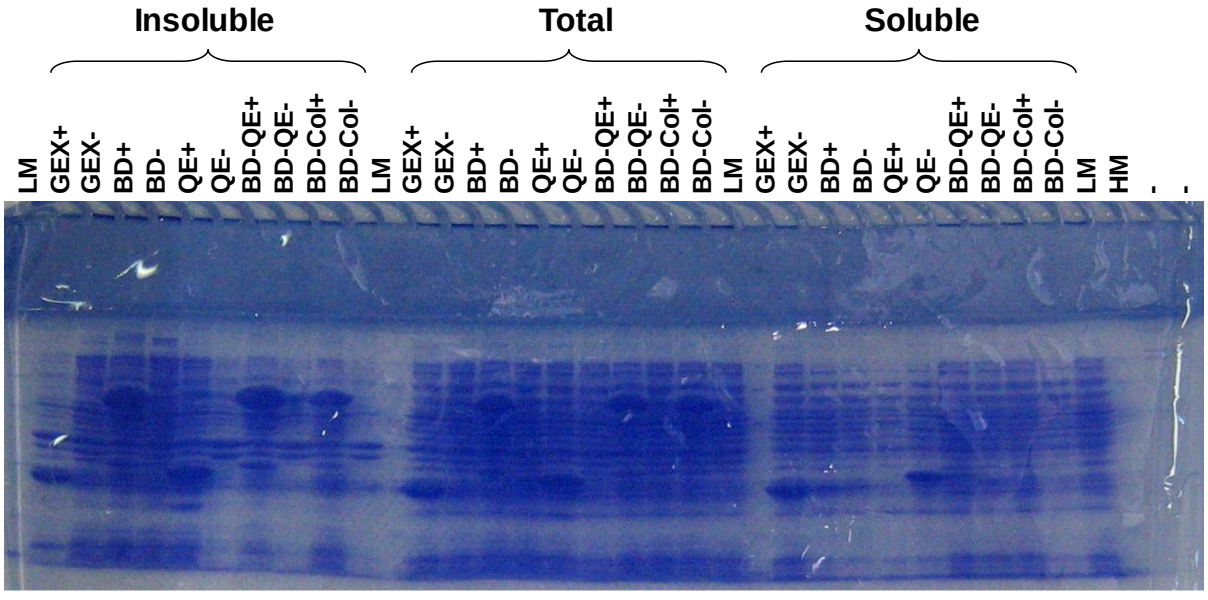


Figure 67. The result of an SDS-PAGE monitoring of the expression and purification of various protein samples. Total stands for the solution, which is obtained by the sonication of the bacterial cells in a certain buffer solution, and it

contains all the proteins present in the cell. The "Insoluble" fractions contain the sample from the pellet after centrifugation of the Total fractions. The "Soluble" fractions represent the supernatant after the centrifugation of the Total fractions.

Looking at the Total fractions, it can be well seen that in each of the lanes marked with + there are some bands, the size, i.e. the intensity of which are much larger than the others. This is not the case in lanes marked with –. The difference between the two types of the samples is that IPTG was added to the LB medium for protein overexpression in bacteria for the samples marked with +. Indeed, these large intensity bands refer to high level of the expression of the target protein, which can not be detected in samples from bacteria incubated without IPTG. The band of the target protein can clearly be identified based on such comparisons and considerations. Furthermore, the molecular size markers also enable the estimation of the size of the protein supporting the above discussion.

To aid the purification of the proteins from such a complex mixture of various proteins, already the protein expression has to be carefully designed. The selection of e.g. the appropriate one out of the several available modified expression vectors that are suitable for recombinant protein expression is the first step in this procedure. The map of e.g. the pGEX-6P-1 expression vector is shown in **Fig. 68**. From the figure it can be learnt that the gene of the protein to be expressed is inserted after the tac promoter. This is an artificially modified lac promoter to increase the efficiency of the transcription. The latter process in pGEX-6P-1 plasmid is regulated by the lac operator, i.e. the transcription can be initiated by adding IPTG to the bacterial culture. IPTG binds to the repressor,

activates the operon, and enables overexpression of the target protein. A common choice for such expression experiment is the *E. coli* BL21(DE3) strain.

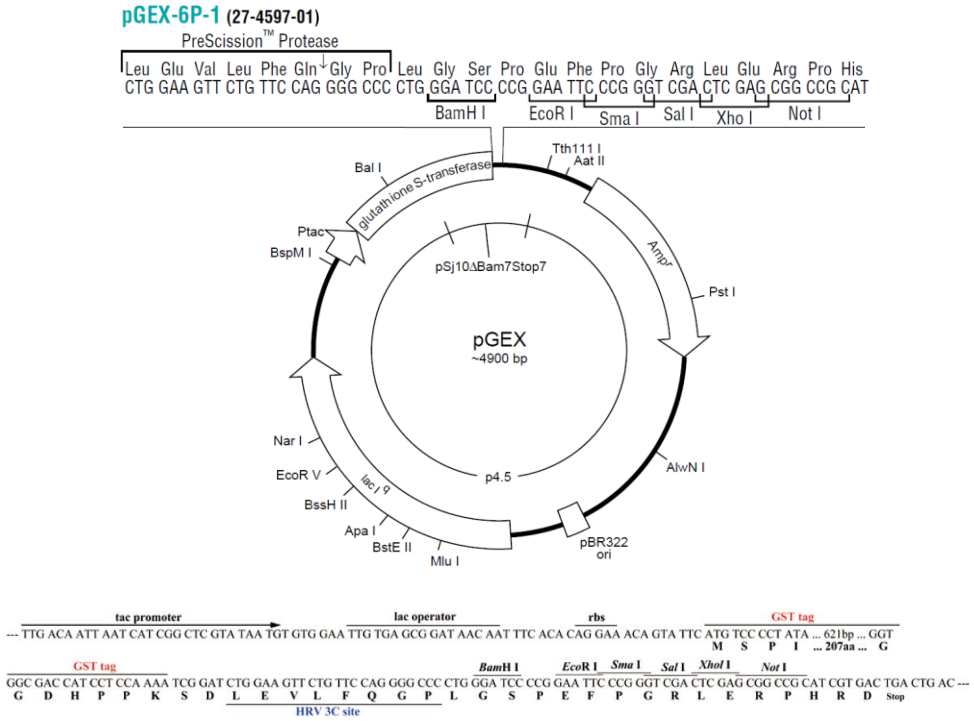


Figure 68. The scheme of pGEX-6P-1 expression vector. The cloning site is enlarged above the vector. In the bottom part of the figure the structure of the operon region is detailed. The ampicillin resistance gene aids to control experiments: if the culture media are supplied with ampicillin, only the cells that have the resistance gene, ie. contain the plasmid can survive.

The pGEX-6P-1 vector contains the gene of the glutathione S-transferase (GST) enzyme between the promoter of the protein expression and the cloning region, without a stop codon. Thus, the result of protein expression will be a GST-fused target protein, consisting of one continuous polypeptide sequence: the N-terminal domain is the GST and the C-terminal is the target protein. There is a

flexible linker between them. For this reason, it is expected that GST does not interfere with the folding and function of the target protein. The advantage to use the GST-tag is that it allows for purification by affinity chromatography. In affinity chromatography, the protein is separated based on specific, reversible interactions established with a ligand that is coupled to a solid chromatographic matrix. The specific interactions most frequently used in affinity chromatography include: enzyme – substrate analogue, e.g. glutathione-S-transferase (GST) – glutathione; antibody – antigen; metal ions – oligo His-sequence. The advantage of this method is the very high specificity and related to this the high sample loading capacity. GST affinity chromatography is using an agarose bead (or sepharose) functionalized with immobilized glutathione – the substrate of the GST enzyme – for protein purification. This procedure results in a high purity product within almost a single step of the purification procedure. After the elution from the resin by reduced glutathione solution, the purified GST-protein can be cleaved with a specific protease, the Human rhinovirus C3 protease (PreScission protease, GE Healthcare). The recognition site of this protease is built in between the GST and the target protein. A short sequence, depending on restriction sites applied for cloning, will remain at the N-terminus of the protein. If the protease is also fused to a GST tag, the cleavage can be carried out already on the washed resin, so that the protease and the GST remain bound to the resin, while the target protein can be eluted from the resin by the desired buffer solution.

Another example of plasmids is pET-21a, that fuses a C-terminal $(His)_6$ sequence to the protein, if the gene does not contain a stop codon before the hexahistidine containing part. Multihistidine fusion tags bind immobilized Ni(II) ions strongly. Thus, His-tagged proteins stay on the resin while the other proteins

are washed away, and then eluted by imidazole solution. A chromatographic column and a Fast Protein Liquid Chromatoraphy (FPLC) instrument are shown in **Fig. 69**. FPLC is a high performance liquid chromatography (HPLC) method developed for the purification of biological samples. This method has the following main characteristics: high loading capacity, biocompatible aqueous buffers, fast flow rates and wide range of stationary phases, such as affinity, gel filtration, ion exchange. The experiment can be automatized by using autosampler, gradient program control and peak collection.

A **B**

Figure 69. A) Immobilized metal ion chromatographic column loaded with nickel(II) ions. **B**) An ÄKTA FPLC explorer system (Amersham Pharmacia Biotech, Sweden) used in the laboratory of the author of this e-book for efficient protein purification.

There is no recommended procedure for removing the C-terminal hexahistidine tag. If this may interfere with the function/structure of the target protein, a specific redesign has to perform, such as it has been done recently in the laboratory of the author of this e-book. As an exercise, search the literature for this procedure.

Plasmids can also be used to express proteins without any affinity tags. However, in this case the protein purification is more challenging. Chromatographic procedures based on the non-specific interactions may be applied, such as the ion exchange chromatography and size-exclusion chromatography (gel filtration).

DNA-binding proteins, such as e.g. the zinc-fingers and the nuclease enzymes are positively charged molecules, complementing the negative charges of the nucleic acids. Thus, they can be purified by cation exchange chromatography. The positively charged solute molecules interact with the negatively charged groups immobilized on the solid matrix. Such cation exchangers possess carboxymethyl groups ("CM"; $–O–CH_2–COO^-$), sulfopropyl groups ("SP", $–O–CH_2–CHOH–CH_2–O–CH_2–CH_2–CH_2SO_3^-$) or methyl sulfonate groups ("S", $–O–CH_2–CHOH–CH_2–O–CH_2–CHOH–CH_2SO_3^-$). In the purification process, first the system is equilibrated with an appropriate buffer, before the sample is loaded onto the column (adsorption). Then an increasing ion gradient and/or pH-change is applied for the desorption of the molecules in the order of their binding affinity to the column. A high resolution can be achieved by optimizing the gradient elution. Ion exchange chromatography is not a specific method thus, even a well resolved peak of the chromatogram can contain more than one type of protein. Sepharose SP Fast Flow column is one of the many

choices for the experiments. Highly crosslinked (6%) 90 μm agarose beads serve as a solid matrix in the Fast Flow ion exchanger columns providing high physical and chemical stability and allowing for high flow rates in a wide pH-range.

During gel filtration the molecules are separated based on different rate of migration in a gel matrix due to their different size. Unlike the previously introduced techniques, the buffer usually has no significant effect on the resolution. Therefore, a wide range of buffers and conditions can be applied – thus, a buffer exchange is also possible using such columns. The separation process can be carried out in the presence of cofactors or denaturing agents and at different temperatures. A long column is used to achieve a high resolution and the sample is injected in a high concentration in low volume. The column is packed with porous spherical particles of gel filtration medium. Molecules diffuse in and out of the pores of the matrix. First the higher molecular weight molecules are eluted, since they can not enter the small pores of the gel particles and therefore, they migrate between them quickly. Smaller molecules, however, move further into the matrix and therefore, stay longer on the column. Since small molecules or ions such as salts that have full access to the pores. Thus, these columns can also be applied for desalting of the protein buffer solution.

For more details of the protein chromatographic methods the reader is directed to the website of the GE Healthcare Life Sciences company: https://www.gelifesciences.com/en/us/solutions/protein-research/knowledge-center/protein-purification-methods

Monitoring questions

- What is the pI value of the proteins.

- List the amino acids encoded by the DNA and characterize their side-chains.

- What is the meaning of SDS-PAGE?

- What is the role of SDS in gel electrophoresis of proteins?

- Which properties of the target protein are the most helpful for its identification on the SDS-PAGE picture?

- What are the fusion proteins and what they are used for?

- How can be fusion recombinant proteins obtained?

- List various types of chromatographic separation methods applied for protein purification purpose!

- Explain briefly the principles of the mentioned chromatographic methods!

Suggested reading

1. Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, James D. Watson: The molecular biology of the cell, Garland Publishing Inc, New York, London, 1989.

2. The Nobel lectures of the Nobel Prize holders listed in this e-book at: https://www.nobelprize.org/

Few suggested articles:

3. Dunkle, J.A., Wang, L., Feldman, M.B., Pulk, A., Chen, V.B., Kapral, G.J., Noeske, J., Richardson, J.S., Blanchard, S.C., Cate, J.H.: Structures of the bacterial ribosome in classical and hybrid states of tRNA binding. Science (2011) 332:981-984

4. Vieira J, Messing J.: The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene (1982) 19(3):259-268.

5. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H.: Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol. (1986) 51 Pt 1:263-273.

6. Yin YW, Steitz TA.: Structural basis for the transition from initiation to elongation transcription in T7 RNA polymerase. Science (2002) 298(5597):1387-95.

Acknowledgements

The author is thankful to all the colleagues for contributing to the friendly atmosphere, which allowe to introduce the biological tools and establish the Artificial Metalloproteins Laboratory at the Department of Inorganic and Analytical Chemistry, University of Szeged.

A great acknowledgement goes to Professor Kyosuke Nagata from University of Tsukuba Japan, who introduced the author into the basics of molecular biology. The author spent a year as a postdoctoral fellow in his laboratory within the frame of a UNESCO/TOKODAI fellowship studying the factors affecting the reproductive cycle of Adenovirus. All the laboratory members supported the author and became not only excellent collaborators, but very good friends, as well.

The enthusiasm of our collaborators on the area of artificial metalloproteins helped the author to continue his research on this new field of Bioinorganic Chemistry. Besides the Japanse laboratory, Antal Kiss and Éva Hunyady-Gulyás (Biological Research Center, Szeged), Hans E.M. Christensen (Danish Technical University, Lyngby), Sine Larsen, Peter W. Thulstrup, Lars B.S. Hemmingsen (Copenhagen University), Søren V. Hoffmann and Nykola Jones (Århus University), Wojciech Bal (Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw), Chris Oostenbrink (BOKU, Vienna), Milan Kožíšek (Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Prague) and their colleagues were the most active participants of this adventure.

The valuable contribution of all the former and present colleagues, postdoctors, PhD students and students to the researches carried out in the Artificial Metalloproteins Laboratory was indispensable to collect the necessary knowledge for a success of the team. These young researchers inspired the author to write the present e-book.

Special thanks goes to Dr. Eszter Németh, former member of the laboratory, for figures taken from her PhD thesis, and to PhD students Heba A.H. Abd Elhameed and Bálint Hajdu for reading and correcting the text.