

## Chapter 2

# Analyzing Biological Structure and Function

This chapter will offer a very brief introduction to biological structure and function. For a more in-depth coverage, please refer to an up-to-date biochemistry or molecular biology book. We have already encountered the biological building blocks of DNA bases and amino acids in chapter 1. This chapter will describe some structural features and the related function of the macromolecular polymers formed from those building blocks. The chapter should also give a very concise overview about the tools of molecular biology which allow the manipulation and analysis of biological function.

### 2.1 DNA structure

DNA is a macromolecule consisting of DNA bases, sugars and phosphates and carries the genetic code. The structure of DNA is depicted in figure 2.1. Two strands of DNA with complementary sequence usually adopt a right-handed double-helical structure ( $\beta$ -helix) with 23.7 Å diameter and 10.5 bases per turn ( $\approx 35$  Å step height). The double helix contains a major and a minor groove which are about 12 and 22 Å wide. DNA melts between 60 and 100 °C, the melting point is higher for GC rich DNA than for AT rich DNA because the GC base pair forms three H-bonds to the two H-bonds between A and T.

In prokaryotes (bacteria and archaea), a single, usually circular DNA molecule carries the complete genetic information. An example is the bacteria E-coli with a genome of  $4.6 \times 10^6$  base pairs containing some 4400 genes. In eukaryotes (animals, plants, fungi), the genetic information is stored on multiple linear DNA strands (chromosomes) which are packed into a cell nucleus. An example is the human genome with  $\approx 3.4 \times 10^9$  base pairs spread over 46 chromosomes (23 pairs) containing 20-25 thousand genes. It may appear curious why the human is ten times larger than that of E-coli if it encodes only five times the number of genes. Indeed, it is now understood that 95% of the human DNA are non-

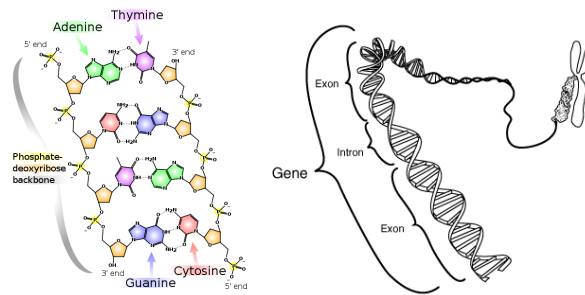


Figure 2.1: Structure of DNA: The four DNA bases adenine, thymine, guanine and cytosine are bound to a sugar phosphate backbone. Two DNA strands with complementary sequence (pairing of A-T and G-C in the H-bonded Watson-Crick base pairs) form a double stranded DNA helix. A gene is a stretch of DNA coding for a protein and may contain regions coding for the protein (exons) and non-coding regions (introns). The complete DNA molecule (chromosome) is further folded, e.g. by interaction with histone proteins in eucariotes. Images adapted from Wikipedia.

coding ("junk-DNA", introns), but is probably responsible for some aspects of gene regulation. The genome size ( ? ) is not directly related to the organism size. A frog genome has a size of some 6500 Mb (1 Mb corresponds to  $10^6$  base pairs with a weight of about 978 ng), the goldfish genome has a size up to 3000 Mb in 50-150 chromosomes.

To create offspring, an organism must copy its own genetic information in a process called DNA replication (2.2). This involves the splitting of the two DNA strands of the DNA double helix, followed by the synthesis of a new corresponding strand for both single-stranded molecules. To facilitate the splitting of the DNA double helix, a protein "helicase" unwinds the twisted double strand and a "topoisomerase" releases the unwinding strain by cutting one strand and ligating it after relaxation of the rotational stress (ligation = formation of the covalent backbone bonds). The DNA polymerase complex then synthesizes the complementary strand for each single stranded DNA molecule.

## 2.2 DNA coding for proteins

The DNA sequence in a gene is coding for the amino acid sequence in proteins. Three DNA bases are needed to code for a single amino acid and are called a codon (see figure 2.3). Thus every single of the 20 amino acids can have a unique DNA code ( $3 \times 4$  distinguishable bases allows for  $4^3 = 64$  unique combinations). There are more codons than amino acids, hence there are several codons for each amino acid and there are three codons coding for the end of a gene ("stop codons"). The availability of several codons for the same amino acid is useful, because it allows for flexibility in the genetic code without forcing changes to the protein sequence. A change in the GC content of the genome, for example,

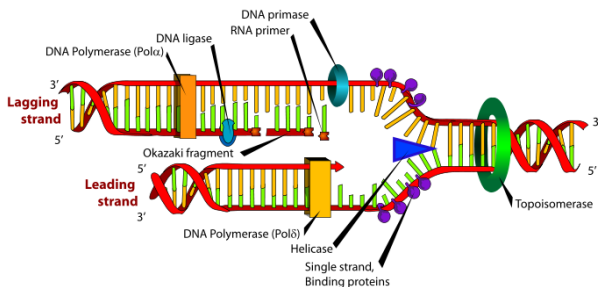


Figure 2.2: The polymerase chain reaction: DNA polymerase extends a piece of DNA along a template DNA strand by catalyzing the formation of the phosphate backbone bonds starting from desoxyribonucleotide-triphosphates (dNTPs) and matching the template strand. The reaction occurs along the 3-5 direction of the template strand and the opposing strand is built in small increments which are subsequently joined by a ligase. PCR can only occur along a single stranded DNA template and requires previous melting of DNA, which is helped by the unwinding of DNA by a helicase. A topoisomerase can cut and reconnect one strand of ds-DNA to relieve the coiling tension due to the action of the helicase. Image from Wikipedia.

can change the melting point of DNA and help thermophilic bacteria to weather high temperatures.

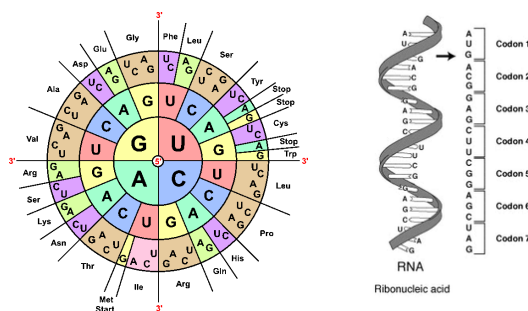


Figure 2.3: The 64 codons of DNA. Image adapted from Wikipedia.

The "reading frame" defines which three bases belong to one codon. A shift of the reading frame by one or two bases will lead to completely different codons and scramble the information on the gene. Hence there is the need for a start codon, which is the same (AUG) in all genes. This codon codes for methionine and therefore all proteins have a methionine as first amino acid.

The expression of a gene into a protein amino acid sequence is a two-step process. First the DNA is "transcribed" by the enzyme RNA polymerase into a single stranded mRNA molecule ("messenger" RNA). The RNA molecule is

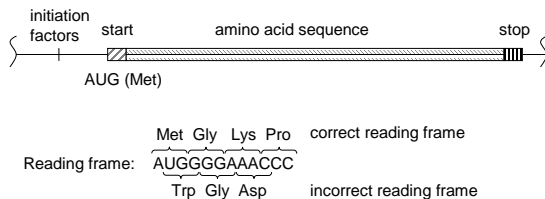


Figure 2.4: Structure of a gene: The coding region of a gene is usually preceded by initiation factors which help to bind transcription factors and thereby initiate transcription and translation. The gene starts with the start codon AUG which defines the correct open reading frame of coding base triplets. The gene ends with one of three stop codons (UGA, UAG, UAA)

very similar to the DNA molecule, but contains an additional hydroxy group in the ribose sugar unit and the DNA base thymine is substituted with the RNA base uracil. The hydroxy group facilitates the hydrolysis of the RNA molecule, therefore the RNA molecule has a much shorter lifetime (minutes) than DNA (centuries). The mRNA may require posttranscriptional processing, e.g. splicing of introns in eucariotic RNA to remove noncoding segments. The spliced introns may perform regulatory functions, but their role is often unclear. The remaining coding gene is translated into the corresponding amino acid sequence by the ribosome, a very large Protein-RNA complex. In a simplistic picture, the ribosome compares the sequence of the next mRNA codon to that of a tRNA (transfer RNA) which carries the correct amino acid residue and attaches the amino acid to the growing chain of the peptide.

## 2.3 Molecular biology tools

### 2.3.1 polymerase chain reaction (PCR)

Parts of the biological machinery for manipulation of DNA have been harnessed and allow the controlled creation, mutation and expression of DNA. DNA polymerase can create copies of a DNA strand starting from a primer (a short sequence of the complementary DNA strand) as shown in Fig. 2.2. This process can be used to amplify natural or synthetic DNA. A mayor advance was the discovery of a heat resistant DNA polymerase which allows the duplication of a DNA strand by repeated melting of double stranded DNA into single stranded DNA followed by duplication of both single stranded DNA strands at lower temperatures. The melting point of DNA is close to 95 °C. Each heat cycle multiplies the amount of DNA and repeated heat-cycling allows strong amplification even from a single DNA molecule. The techniques are mature

enough that scientists hope to analyze the DNA of extinct animals, such as the mammoth, from tiny amounts of highly degraded DNA.

### 2.3.2 Sequencing

To recognize and manipulate meaningful DNA sequences, it is necessary to read the genomic information of cells and virii. A number of methods for reading the DNA sequence, i.e. "sequencing" have been developed in the history of molecular biology. The most efficient method currently in use is the creation of truncated DNA strands through PCR (see above). PCR requires deoxynucleotide-triphosphates (dNTP) to elongate a DNA strand according to a template strand. If a dideoxynucleotide-triphosphate (ddNTP) is incorporated, the strand cannot be extended due to lack of the necessary hydroxy group on the ribose and the length of the truncated strand yields information about the position of the ddNTP. Four separate reactions with ddATP, ddTTP, ddGTP and ddCTP can be used to read out the complete sequence (see figure 2.5). The length of the truncated DNA strands is determined by gel electrophoresis, where the molecules are dragged through a polymer gel by an electrical field: the molecules travel a distance inversely related to their size. The DNA was typically labelled with radioactive markers (e.g.  $^{32}\text{S}$  or  $^{35}\text{S}$ ) and the DNA strand positions could be read out by exposing a photographic plate to the radiation.

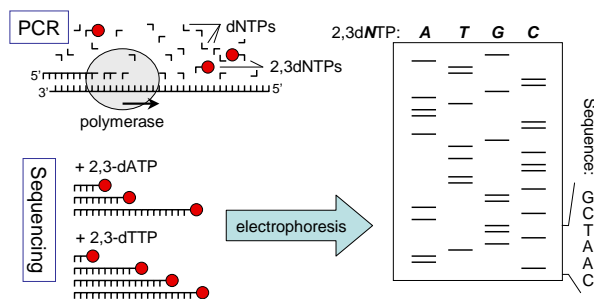


Figure 2.5: The Sanger method for gene sequencing: single stranded DNA is incubated with a short primer ( $\approx$  20 bases) and the corresponding strand is synthesized with PCR in four separate reactions, each containing a small amount of ddNTP of one base. The reaction is truncated when a ddNTP is incorporated and the DNA strand length corresponds to the position of that ddNTP and to the position of the corresponding DNA base in the template strand. The length of the truncated DNA strands is determined by gel electrophoresis, where the molecules are dragged through a polymer gel by an electrical field.

Modern methods use fluorescing dye terminator ddNTPs which allow the distinction of the four bases by their fluorescence properties (see Fig. 2.6). In this approach, only a single PCR reaction and a single trace on the Gel is necessary, greatly increasing the efficiency of the procedure.

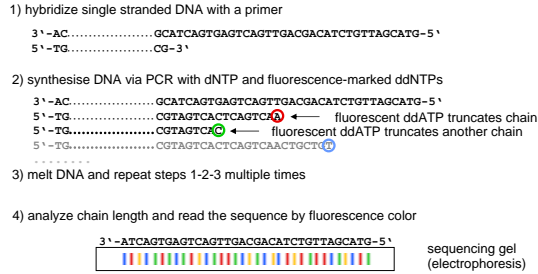


Figure 2.6: Modern Sanger method for gene sequencing: The complementary strand for single stranded DNA is synthesized with PCR as described in Fig. 2.5. The synthesized strand is terminated by a ddNTP which carries a fluorophore. single stranded DNA is incubated with a short primer ( $\approx$  20 bases) and the corresponding strand is synthesized with PCR in four separate reactions, each containing a small amount of ddNTP of one base. The reaction is truncated when a ddNTP is incorporated and the DNA strand length corresponds to the position of that ddNTP and to the position of the corresponding DNA base in the template strand. The length of the truncated positions can be read out after gel electrophoresis. To read out the positions on the gel, the DNA is labelled with radioactive markers (e.g.  $^{32}\text{S}$  or  $^{35}\text{S}$ ), or with fluorescence markers.

Until the mid-1990s, gene sequencing was a very orderly business and molecular biologists would first create a restriction map (see section 2.3.3 below) for a genome or chromosome of interest, and later sequence the restriction fragments. For the determination of complete large genomes (e.g. the human genome project), a faster but less ordered approach was developed: shotgun sequencing. In this approach the DNA is cut into pieces suitable for sequencing without creating a restriction map. The DNA is sequenced and computers search for overlapping regions between the fragments to puzzle together the complete genome. This "small brain, big computer" approach removes the need for multiple restriction reactions and therefore requires less human work and input. Via developments like this, DNA sequencing is nowadays highly automated and standardized.

### 2.3.3 Cutting and connecting

Restriction enzymes can be used to cut DNA at selected sequences, e.g. the enzyme Eco-R1 will cut double stranded DNA containing the GAATTC as shown in Fig. 2.7. A restriction map of a large piece of DNA can be created by using different restriction enzymes and yields smaller and more manageable DNA strands for further DNA sequencing. To get an impression about the large number and variety of commercially available restriction enzymes, visit a restriction map website, e.g. <http://www.restrictionmapper.org>. Ligase can be used to covalently connect two pieces of DNA, and can be used to include

modified DNA into the genome of an organism. Therefore it is possible to modify the genome in a controlled manner without need for a complete de-novo synthesis.

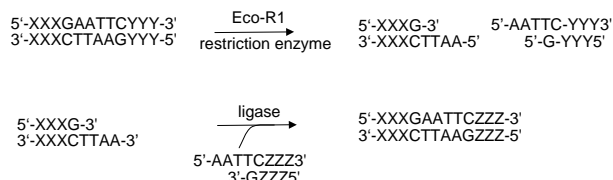


Figure 2.7: Restriction enzymes, such as Eco-R1 can cut DNA at well defined sequences. Ligase can be used to join two DNA molecules with complementary overlapping sequences or blunt ends, e.g. to attach a synthesized sequence ZZZ into the DNA cut by Eco-R1.

The description above is a bit simplistic, because DNA requires a living cell to perform its function. It is not feasible to remove the complete DNA from a cell, e.g. E-coli, modify it and reinsert it into the cell. Instead, small autonomous pieces of DNA can be used, namely plasmids and virii (an example is shown in Fig. 2.8). Plasmids are circular pieces of DNA which carry several genes and all necessary noncoding sequences to allow gene expression. Examples for the latter are the stop codon, but also promoter regions before the gene which facilitate binding of the RNA polymerase and regions which initiate DNA replication. To allow selection of cells which contain the plasmid (transfected cells) against cells which do not, additional antibiotic resistance genes are often placed into the plasmid to allow antibiotic poisoning of non-transfected cells. Placing specific promoters onto the plasmid may allow to switch on/off the gene expression in a controlled manner.

### 2.3.4 Cloning

One common goal in molecular biology is the expression of a known gene in enhanced quantities or in a different organism. The gene is cloned (copied) by suitable digestion of the parent genome and PCR amplification of the gene, possibly including suitable recognition sequences for ligation. The gene is then ligated into a suitable vector or virus which can be used to transfect a cell. A famous early example was the cloning of human insulin genes into bacterial cells to produce insulin for the treatment of diabetes. The human insulin produced in this fashion replaced pig insulin which often created immune reactions in the patients. Monsanto cloned pesticide resistance genes into farmed plants to allow efficient and selective pesticide use, but the resulting genetically modified plants

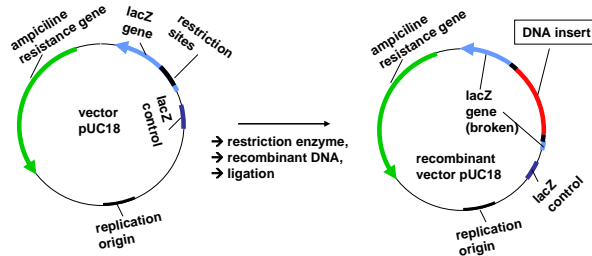


Figure 2.8: Insertion of DNA into a plasmid. The plasmid is cut by a restriction enzyme and ligated with recombinant DNA. The plasmid can then be used to transfect cells and induce expression of the inserted gene. To facilitate selection of the transfected cells, the plasmid carries an antibiotic resistance and a *lacZ* gene. *LacZ* hydrolyses X-gal into a blue product, hence when the gene is destroyed by the DNA insert the cell culture will turn from blue to white.

are not welcome everywhere. In this respect, it is useful to consider that genes also migrate naturally (without genetic manipulation) within species and even beyond species boundaries. Even the natural process of breeding drastically affects the natural selection of gene migration. In general, however, the natural selection (resulting in "wild type" organisms) ascertains that artificially bred, or genetically modified organisms can only survive through extensive human care because they carry the burden of additional genes which do not advance the general survivability.

### 2.3.5 Genetic libraries

To allow the identification, characterization, and cloning of genes, it is useful to create a genetic library for the organism of interest. This is done by partial digestion of the genome and the indiscriminate cloning of all fragments into a fast-growing host (e.g. *E. coli* bacteria). This library can be used to search for specific sequences via Southern blotting (see chapter 2.3.6 below), to identify RNA or protein function upon expression, or to study binding properties of expressed proteins.

### 2.3.6 Southern, Northern and Western Blots

Edwin Southern developed a technique to identify and isolate DNA strands of a selected sequence from a gene library. He found that after DNA separation via electrophoresis in a gel, he could transfer the DNA fragments onto blotting paper. The binding of DNA to the blotting paper is strong enough to allow further chemical treatment and he investigated the binding to other DNA by exposing the paper to a corresponding solution. After washing, the

# Bibliography

- [1] "Biochemistry", Donald Voeth, Judith G. Voeth (edt.), John Wiley and Sons inc., New York 1995.
- [2] "Biochemistry and Molecular Biology", Keith Wilson, John Walker (etd.), Cambridge University Press 2005.
- [3] Online Animal Genome Size Database, Release 2.0, <http://www.genomesize.com>.