

# Introduction and motivation

There are two different aspects of what is called *Bioinfomatik* in German literature:

- ▶ **Computational biology:** Here one aims to answer questions relevant to biological research by using electronic data processing e.g. in order to process huge sets of biological data.
- ▶ **Bioinformatics:** Here we address questions from computer science, mostly of an algorithmic nature, motivated by applications from computational biology.

A third aspect of relevance often is disregarded: Nature as role model for solving algorithmic problems. Examples are

- ▶ evolutionary algorithms,
- ▶ ant colonies,
- ▶ DNA-computing.

In this lecture we deal with bioinformatics mainly focusing on questions arising from molecular biology. The lecture *Nature Inspired Computing* addresses the before mentioned aspect.

## Content

- ▶ **String Algorithms**

Since many objects from molecular biology are modeled as words (DNA-sequences, amino acid sequences,...), efficient string-matching algorithms, i.e. algorithms for finding all occurrences of one string (pattern) within another one (text), are a fundamental component of many software tools in this area. Therefore, we will pay attention to efficient solutions to this problem based on preprocessing the pattern and on the data structure *suffix-tree*, which allows for efficient searches by preprocessing the text.



## ► DNA Sequencing

Sequencing a DNA molecule, i.e. determining its sequence of nucleotides (bases), is a major task of molecular biology. Since lab techniques used for this task are subject to some restrictions, sequencing cannot be done for the entire molecule in a single step. Instead one has to divide the molecule into parts, which then are sequenced. Unfortunately, the relative order of fragments gets lost during this process and an algorithmic reconstruction of the original molecule becomes necessary.

This lecture will address two different approaches:

- *Physical mapping*, which has been used for the human genom project, and
- *shotgun sequencing*, used by Celera Genomics.

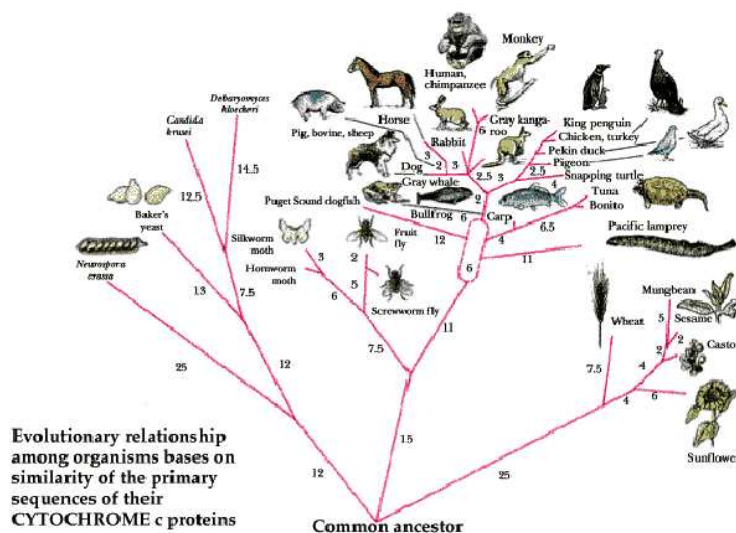
## ► Signal Recognition

By sequencing a DNA molecule we only get knowledge of its sequence of nucleotides (bases) but nothing is known about the function (semantics) of this sequence or subsequences. *Signal recognition* thus aims for detecting functional aspects of the molecule like e.g. binding place of proteins or genes within a DNA sequence. In this lecture mostly statistical approaches will be addressed; we will study the relative frequency of substrings of a DNA sequence and so called Hidden-Markov-Models which have a wide range of applications in computational molecular biology.

► **Phylogenetic Trees (Phylogenies)** (Part 2 of this course)

Phylogenetic trees are used for representing evolutionary interrelationships among various species or other entities that are believed to have a common ancestor. However, the structure of the tree is unknown in advance and has to be derived from information about the species available. This can be done based on some notion of distance e.g. of homologous genes of the different species where the tree then is structured such that species with the most similar genotype are placed closest to each other.

This lecture covers several strategies for constructing a phylogenetic tree using different kinds of information.



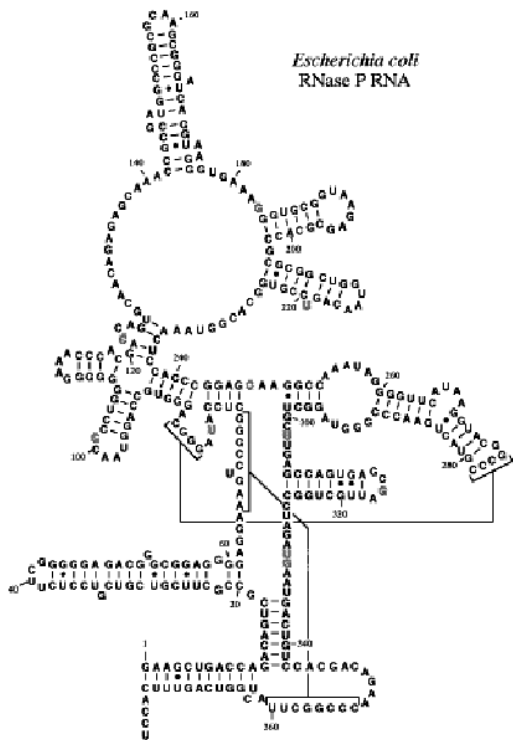
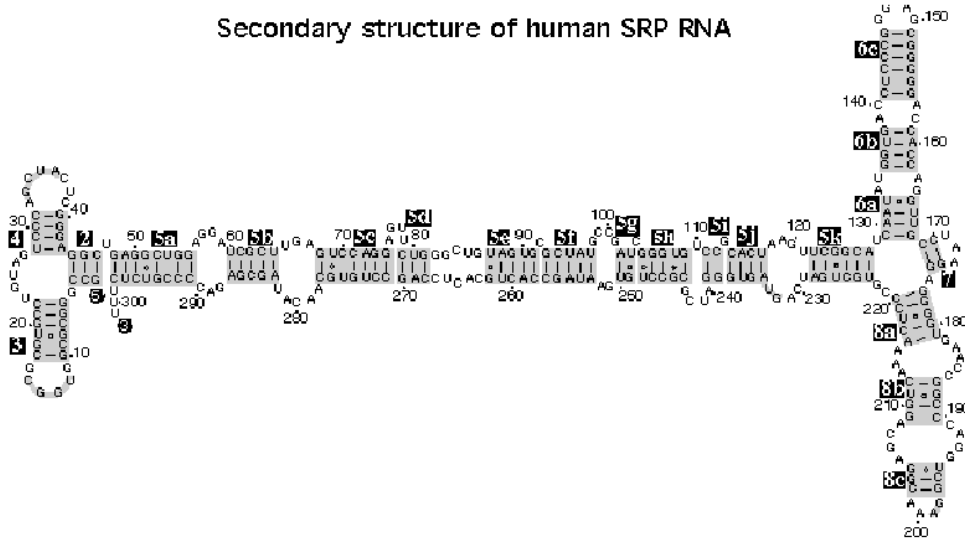
► **RNA Secondary Structure Prediction** (Part 2 of this course)

From a chemical point of view RNA molecules are very similar to DNA molecules with the major difference that RNA is single stranded. The same interactions which force the double stranded DNA into its helical structure allow for bonding between different regions of a single RNA molecule. As a consequence, RNA folds into a 3D structure where different conformations (topologies) are possible. The function of numerous kinds of RNA molecules is determined by its topology. Therefore we aim to predict the spatial structure of an RNA molecule algorithmically from its known sequence.

In order to simplify the problem one considers the so called *secondary structure* as a first approximation of the molecule's topology by allowing only bonds which let the self-folding remain planar. The lecture will present two different approaches of practical importance for predicting the secondary structure, namely

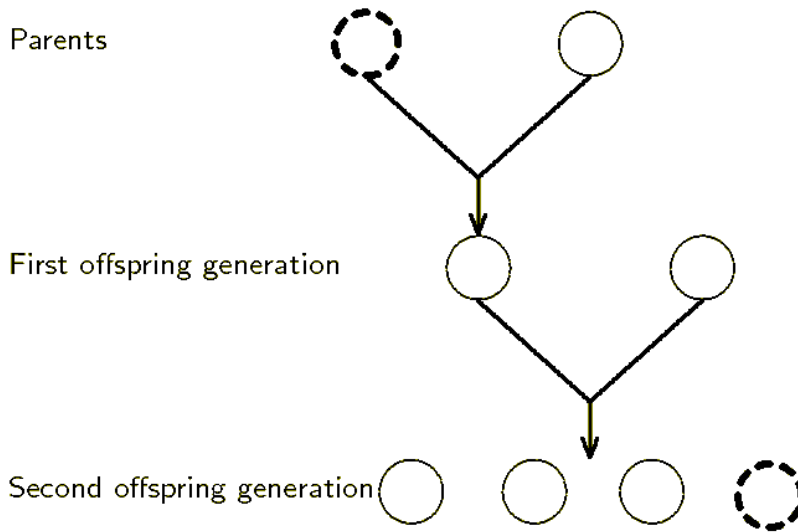
- ▶ the Zuker-algorithm, and
- ▶ stochastic context-free grammars.

Secondary structure of human SRP RNA

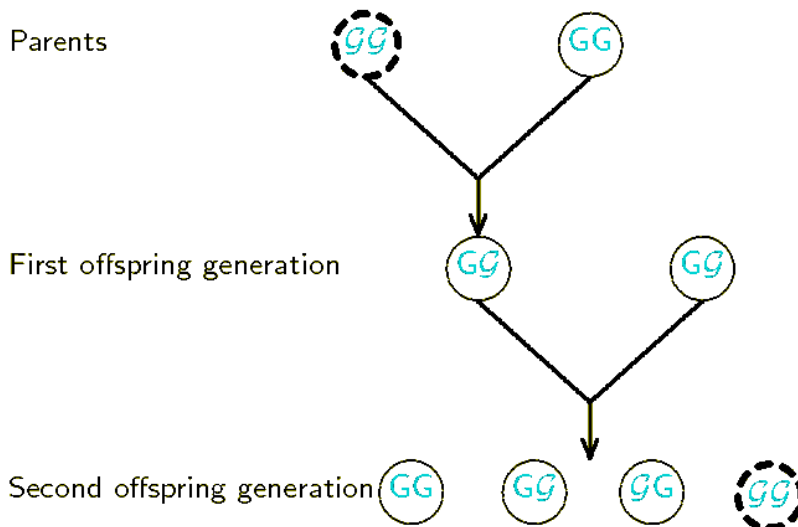


# Primer on Molecular Genetics

**Mendel's Genetic:** Gregor Mendel in the 19th century performed the following experiment in order to investigate inheritance:



**Explanation:** Every offspring (a pea in case of this experiment) inherits one allele (here  $g$  resp.  $G$ ) from each parent for each trait (an allele is one possible version of a gene).



## Details:

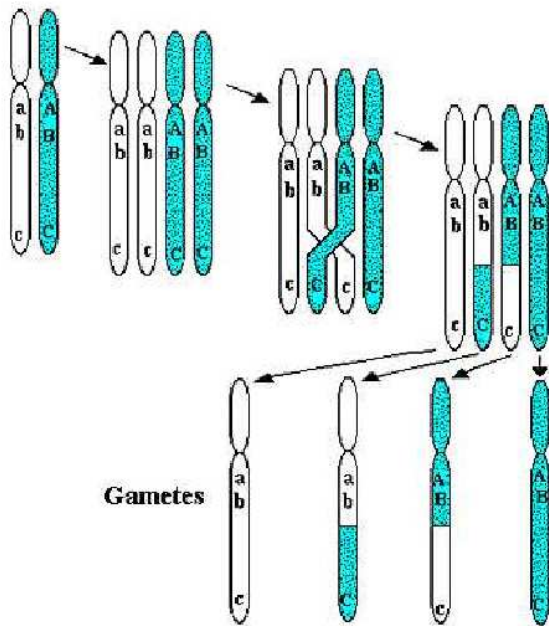
- ▶ **Genotype** (the specific allelic combination for a certain gene or set of genes ) and **phenotype** (literally means “the form that is shown”; it is the outward, physical appearance of a particular trait);
- ▶ **homozygote** (e.g.  $GG$ ) and **heterozygote** (e.g.  $Gg$ );
- ▶ **dominant** (here  $G$ ) resp. **recessive** alleles (here  $g$ );
- ▶ **intermediate phenotype/incomplete dominance** (alleles for the colors red and white yield pink);
- ▶ **co-dominance** where neither phenotype is dominant, instead, the individual expresses both phenotypes.

Implications of the experiments: (Mendel's rules of inheritance)

1. If two individuals of one species differing in only one trait (for which both are homozygous) are mixed then all individuals of the first offspring generation are equal.
2. When mixing individual of the first offspring generation, the second offspring generation does not remain uniform but separates according to the ratio 1:3 (dominant/recessive) resp. 1 : 2 : 1 (incomplete dominance).
3. Mixing individuals of one species which homozygosity differ according to several traits, rules 1 and 2 hold for each single trait. Besides combinations of traits existing in the parent generation, the second offspring generation possesses new combinations.

**Remark:** The set of all genes of an organism is called *genom*, which may be distributed across several *chromosomes*. The third rule holds only for traits of genes located on different chromosomes resp. of genes residing on a single chromosome but with enough distance such that so-called *crossing-over-mutations* are sufficiently probable.

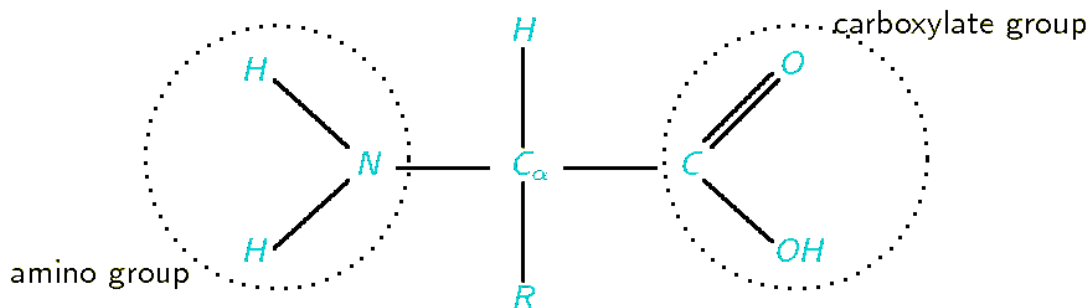




**Crossing-over and recombination during meiosis**

## Proteins:

Proteins are a kind of molecule of high importance to creatures. They work as enzymes to catalyze biochemical reactions, are important in cell signaling and serve as building material. Each protein is made of *amino acids* arranged in a linear chain and joined together by peptide bonds. The chemical structure of a single amino acid is pictured below:



Here  $R$  is the so called side chain, in which the various amino acids differ.



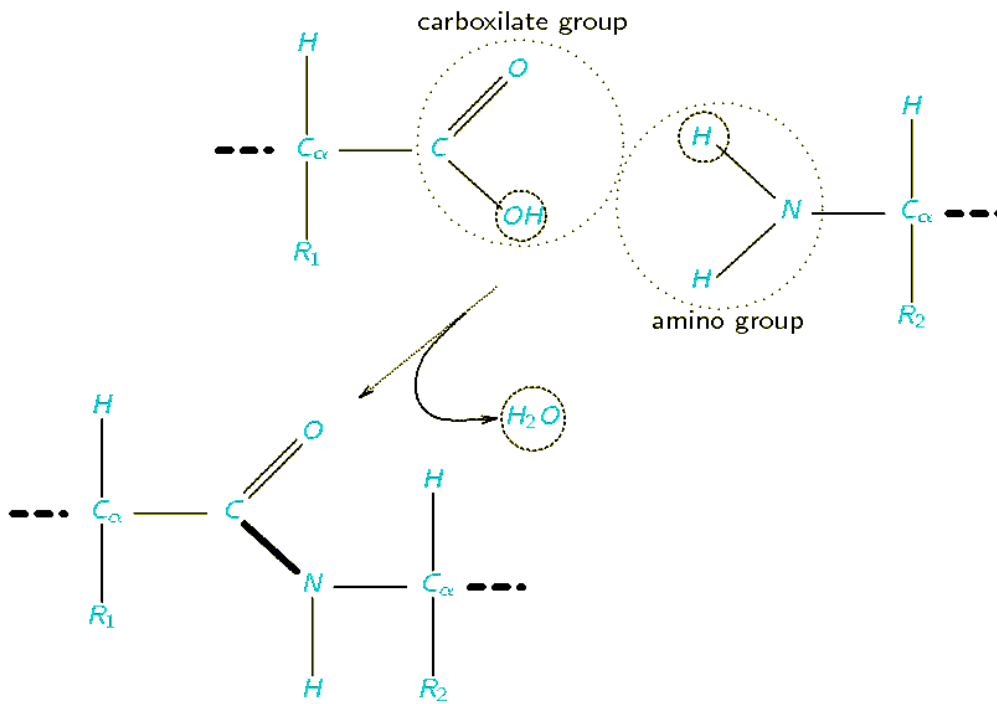
There are 20 different amino acids which occur as building blocks of proteins; the table to the right lists their names and abbreviations:

amino acid	abbreviation	code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

As already mentioned amino acids may combine by a chemical reaction between the amino group of one and the carboxylate group of the other acid ( $\rightsquigarrow$  peptide bond).

$\Rightarrow$  Several amino acids may build a long chain (**polypeptide**) whose backbone is build by the peptide bonds between the different amino and carboxylate groups and the carbon atom  $C_{\alpha}$ .

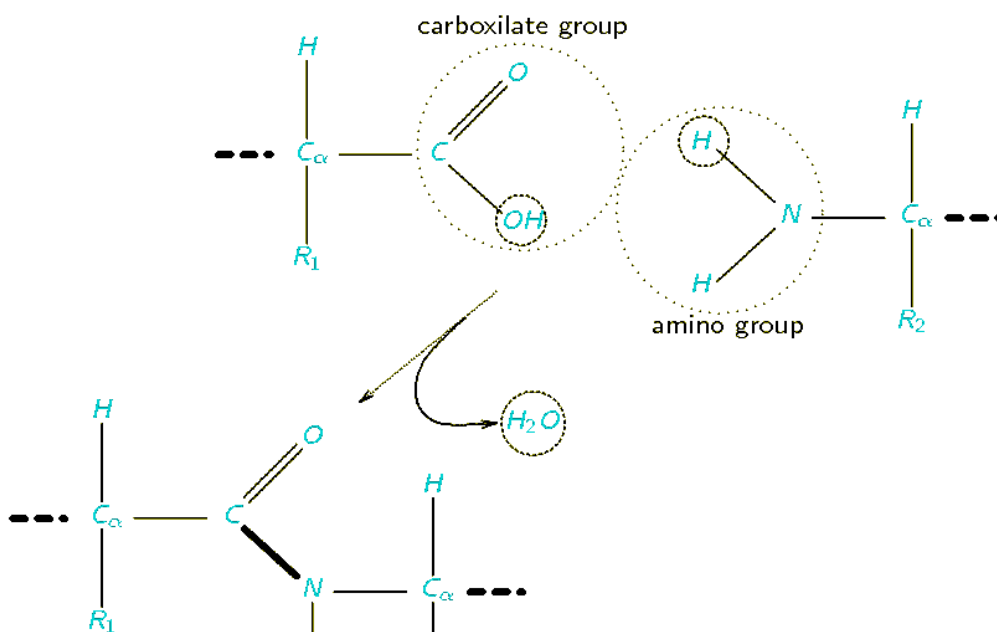
Our image of a protein: polypeptide-chain.



As already mentioned amino acids may combine by a chemical reaction between the amino group of one and the carboxylate group of the other acid ( $\rightsquigarrow$  peptide bond).

$\Rightarrow$  Several amino acids may build a long chain (polypeptide) whose backbone is build by the peptide bonds between the different amino and carboxylate groups and the carbon atom  $C_{\alpha}$ .

Our image of a protein: polypeptide-chain.





As already mentioned amino acids may combine by a chemical reaction between the amino group of one and the carboxylate group of the other acid ( $\rightsquigarrow$  peptide bond).

$\Rightarrow$  Several amino acids may build a long chain (**polypeptide**) whose backbone is build by the peptide bonds between the different amino and carboxylate groups and the carbon atom  $C_\alpha$ .

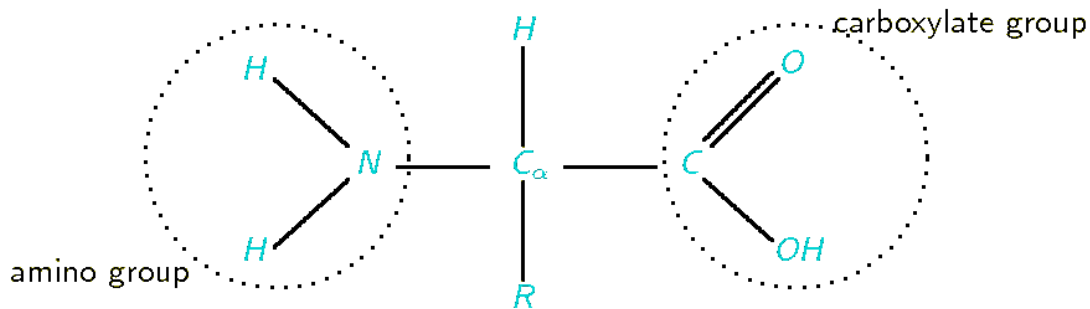
Our image of a protein: polypeptide-chain.

There are 20 different amino acids which occur as building blocks of proteins; the table to the right lists their names and abbreviations:

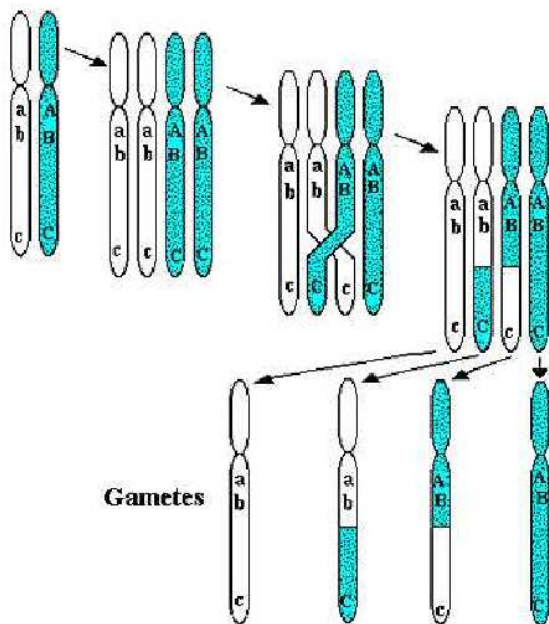
amino acid	abbreviation	code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

# Proteins:

Proteins are a kind of molecule of high importance to creatures. They work as enzymes to catalyze biochemical reactions, are important in cell signaling and serve as building material. Each protein is made of *amino acids* arranged in a linear chain and joined together by peptide bonds. The chemical structure of a single amino acid is pictured below:



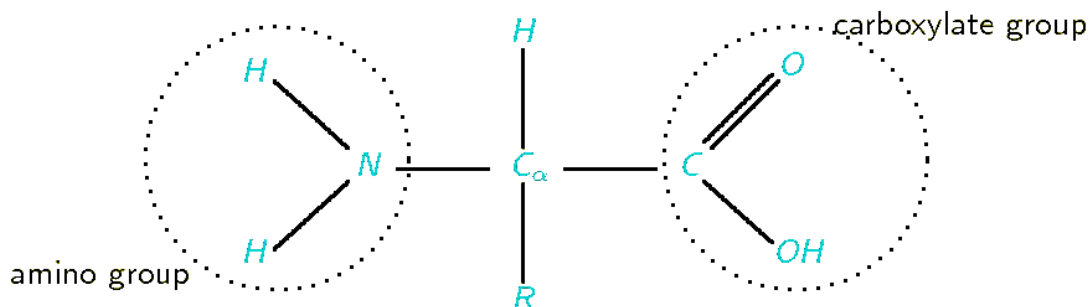
Here  $R$  is the so called side chain, in which the various amino acids differ.



**Crossing-over and recombination during meiosis**

## Proteins:

Proteins are a kind of molecule of high importance to creatures. They work as enzymes to catalyze biochemical reactions, are important in cell signaling and serve as building material. Each protein is made of *amino acids* arranged in a linear chain and joined together by peptide bonds. The chemical structure of a single amino acid is pictured below:



Here  $R$  is the so called side chain, in which the various amino acids differ.

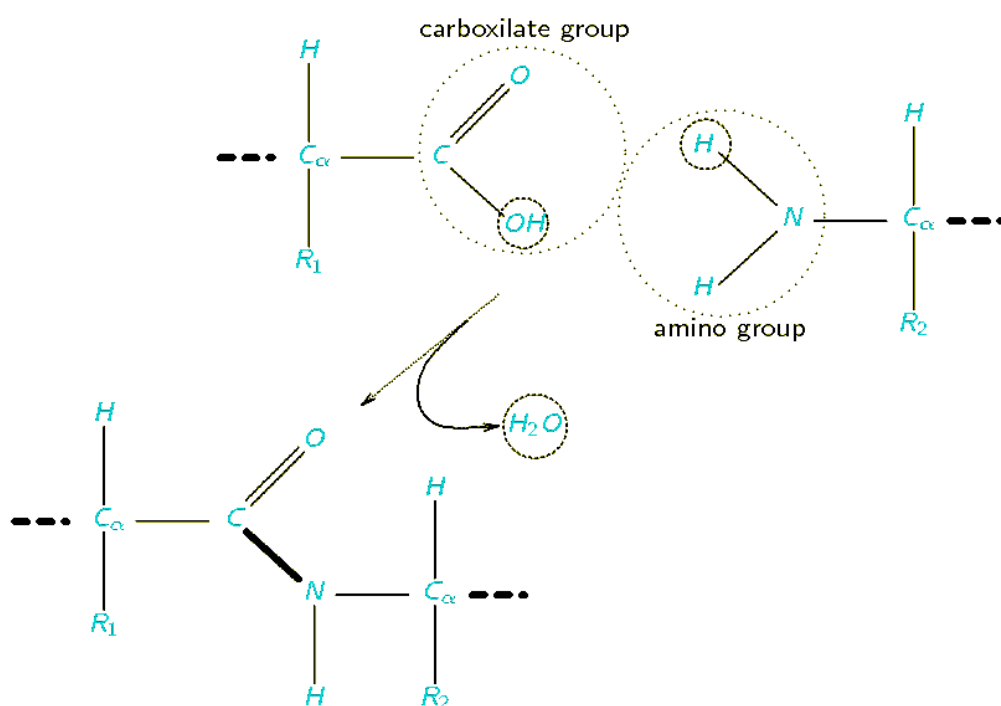
There are 20 different amino acids which occur as building blocks of proteins; the table to the right lists their names and abbreviations:

amino acid	abbreviation	code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

As already mentioned amino acids may combine by a chemical reaction between the amino group of one and the carboxylate group of the other acid ( $\rightsquigarrow$  peptide bond).

$\Rightarrow$  Several amino acids may build a long chain (polypeptide) whose backbone is build by the peptide bonds between the different amino and carboxylate groups and the carbon atom  $C_{\alpha}$ .

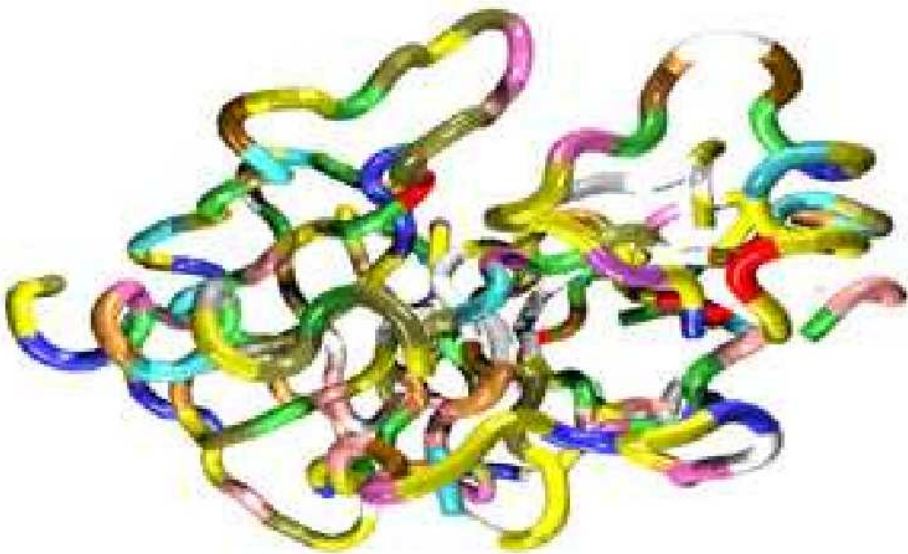
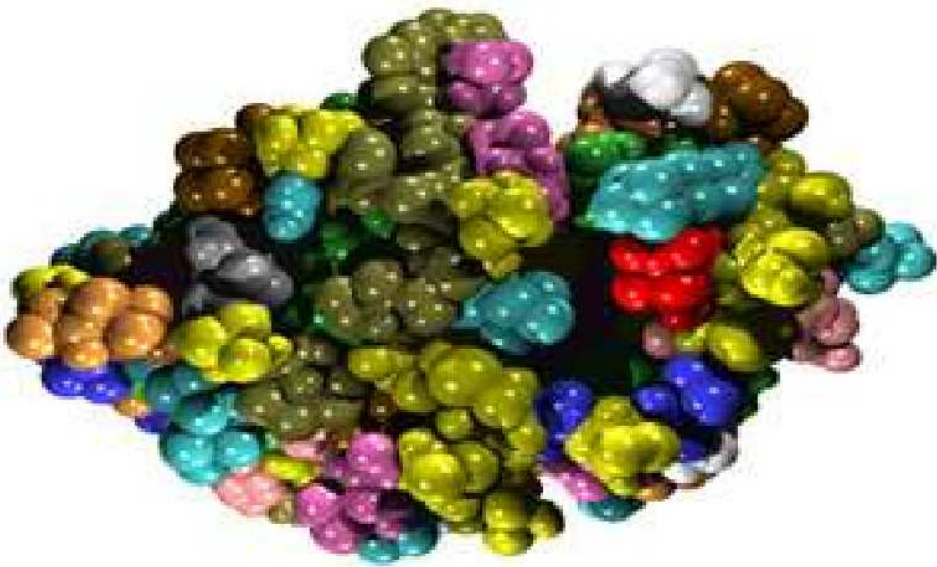
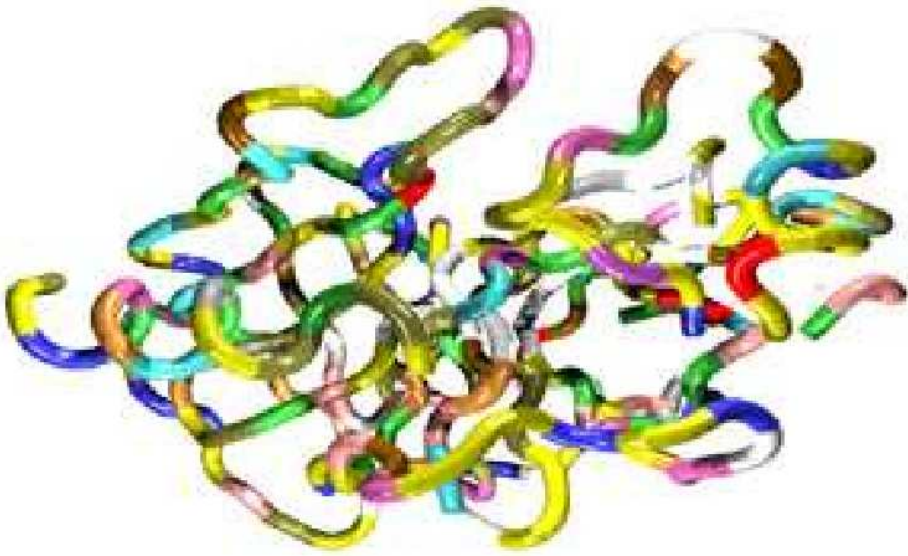
Our image of a protein: polypeptide-chain.



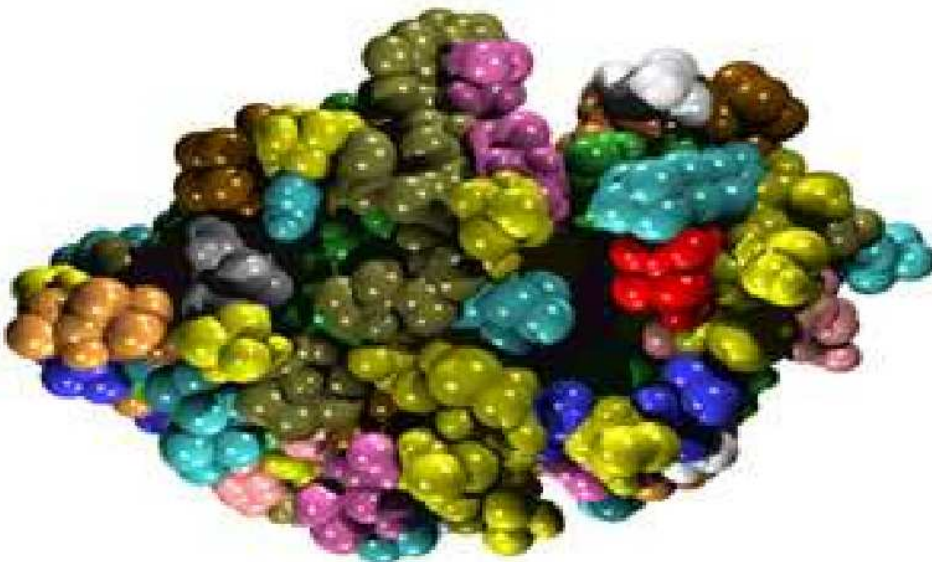
For many applications it will be sufficient, to represent this chain as a word over the alphabet

$\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ .

**However:** (see next slides).



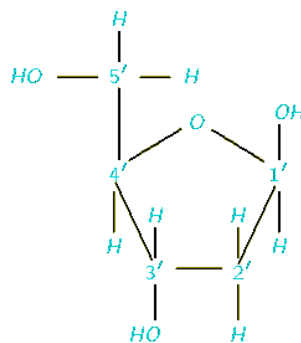
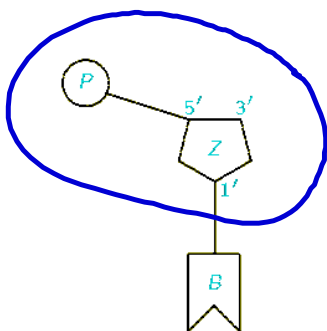




## Nucleic acids:

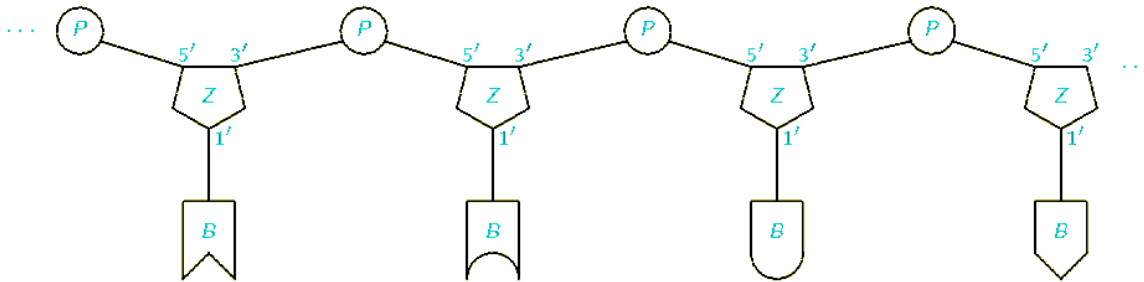
For all species nucleic acids are used to store the genome and thus they provide the structural design (building plan) of the proteins. Furthermore, they are used to transfer genetic information between generations. Additionally, they play an important rôle for the biosynthesis of proteins.

Nucleic acids are build of so called nucleotides which are composed of a sugar *Z*, a phosphate group *P* and a base *B*.



Like the amino acids in proteins, nucleotides in nucleic acids build long chains. This is done by a reaction between the phosphate group of one and the 3'-end of the sugar of the other nucleotide.

If this chain is used for encoding, it has to be read starting with the free 5'-end towards the free 3'-end.

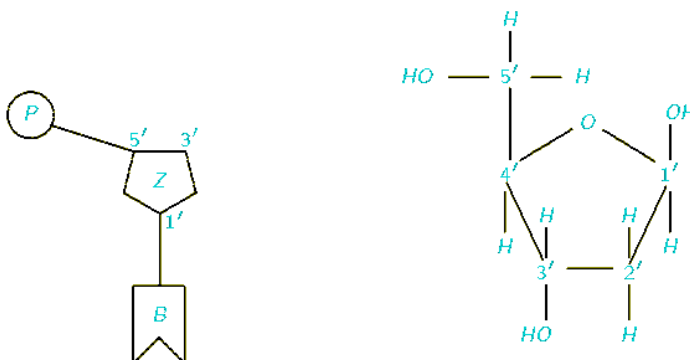


Primary structure: Word over the alphabet  $\Sigma$ .

## Nucleic acids:

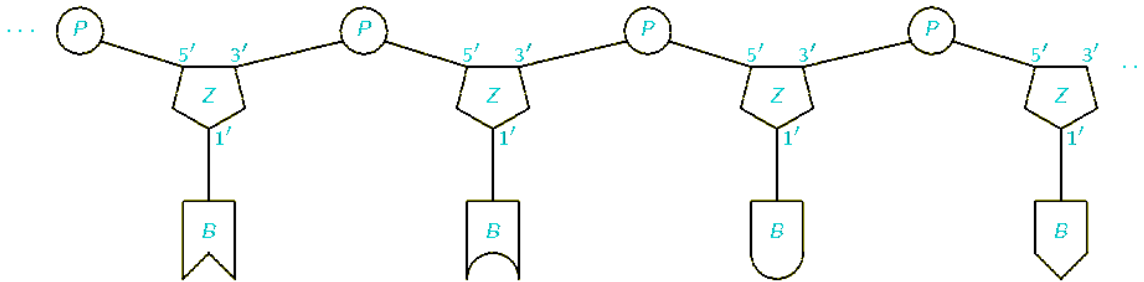
For all species nucleic acids are used to store the genome and thus they provide the structural design (building plan) of the proteins. Furthermore, they are used to transfer genetic information between generations. Additionally, they play an important rôle for the biosynthesis of proteins.

Nucleic acids are build of so called nucleotides which are composed of a sugar  $Z$ , a phosphate group  $P$  and a base  $B$ .



Like the amino acids in proteins, nucleotides in nucleic acids build long chains. This is done by a reaction between the phosphate group of one and the 3'-end of the sugar of the other nucleotide.

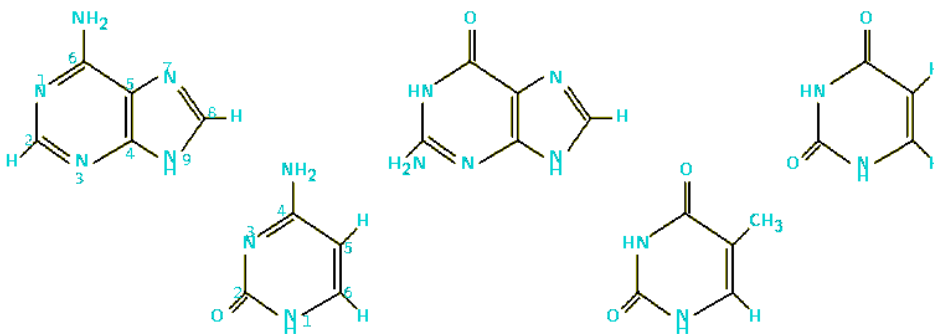
If this chain is used for encoding, it has to be read starting with the free 5'-end towards the free 3'-end.



**Primary structure:** Word over the alphabet  $\Sigma$ .

We distinguish two types of nucleic acids:

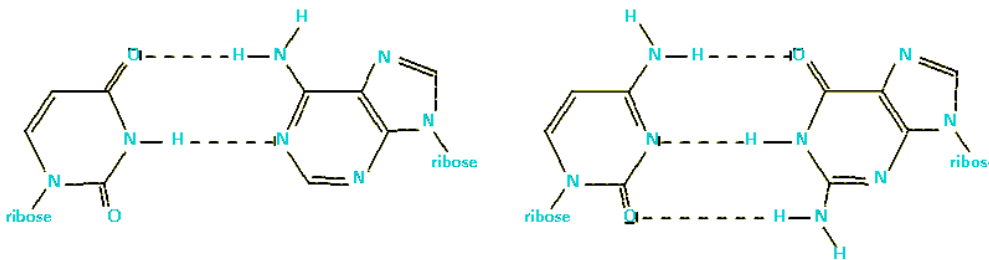
- ▶ **Deoxyribonucleic acid** (DNA for short), where the sugar is a deoxyribose and adenine, cytosine, guanine and thymine occur as bases ( $\Sigma = \{A, C, G, T\}$ );
- ▶ **Ribonucleic acid** (RNA for short), where the sugar is a ribose and uracil replaces thymine ( $\Sigma = \{A, C, G, U\}$ ).



## Base pairing:

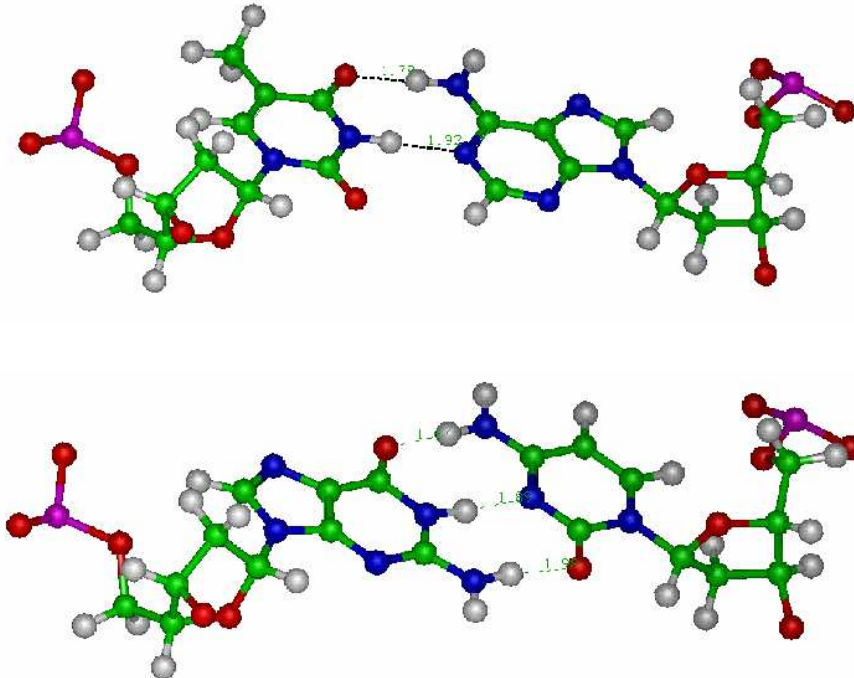
*C* and *G* as well as *A* mit *T* (resp. *U* in case of RNA) may form a hydrogen bond with each other.

Example from the RNA world: (*U* and *A* on the left, *C* and *G* on the right)



For simplifying the presentation, hydrogen atoms at the free places of the rings were omitted.

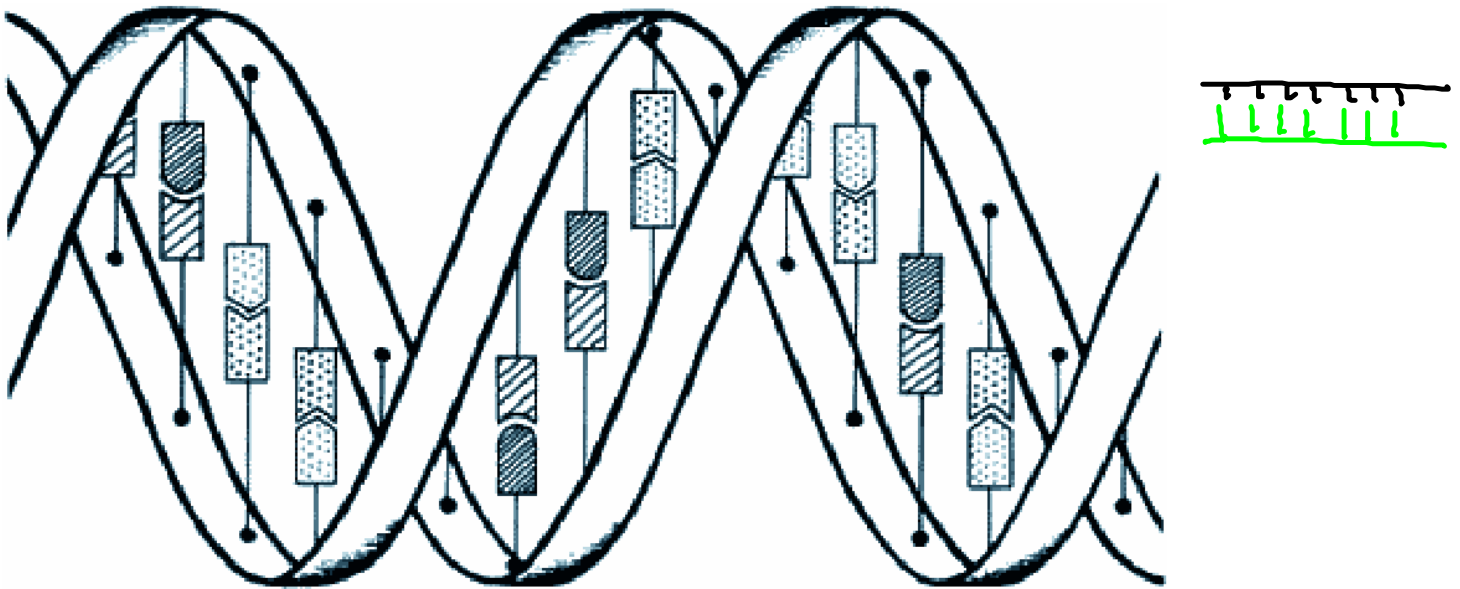
**DNA world** (*T* and *A* on top, *G* and *C* bottom):



## DNA:

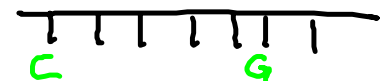
- ▶ We always observe two complementary strands (chains of nucleotides).
- ▶ If position  $i$  (direction  $5' - 3'$ ) of one strand consists of base  $x \in \{A, C, G, T\}$ , then the other strand has the base complementary to  $x$  at position  $i$  (direction  $3' - 5'$ ); a double stranded molecule results from the hydrogen bonds between the complementary bases.
- ▶ This *ladder-like* structure is twisted by effects which will not be discussed here; a double helical structure results.

### Illustration:



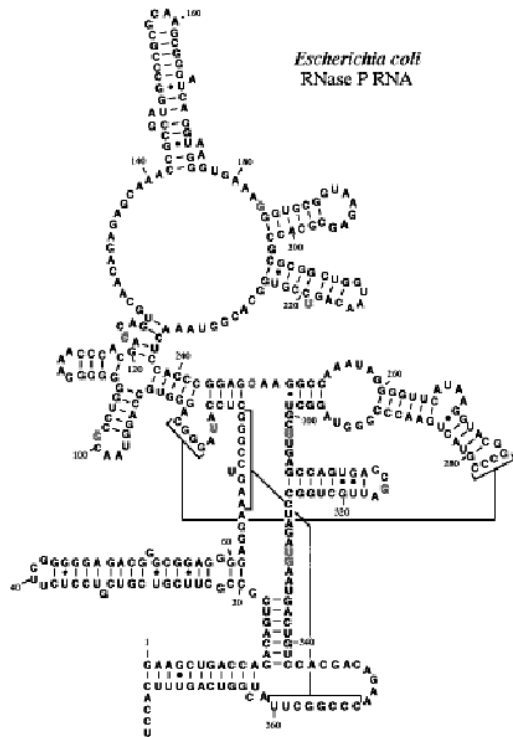
## RNA:

- ▶ RNA molecules are single stranded, however
- ▶ some kinds of RNA are folded with themselves by means of hydrogen bonds between complementary nucleotides of the same strand.
- ▶ The subsequences involved need not be the exact complement of each other; by means of bulges and loops parts which find no counterpart remain unpaired.



⇒ molecule gets a fixed spatial structure and may act e.g. as enzyme.

## Illustration:



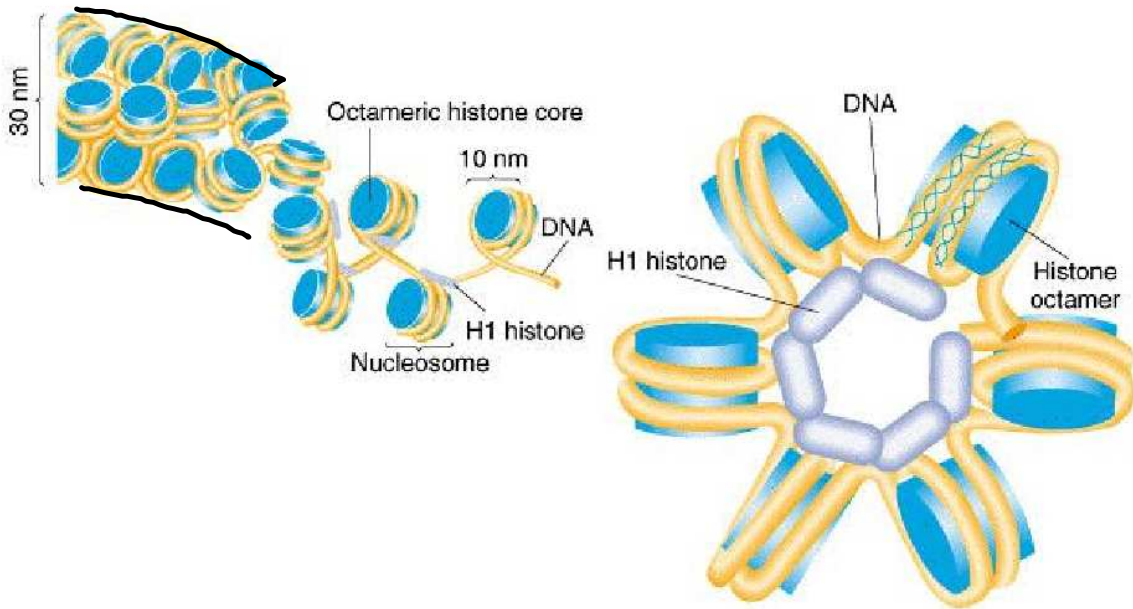
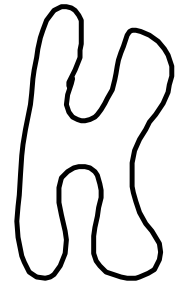
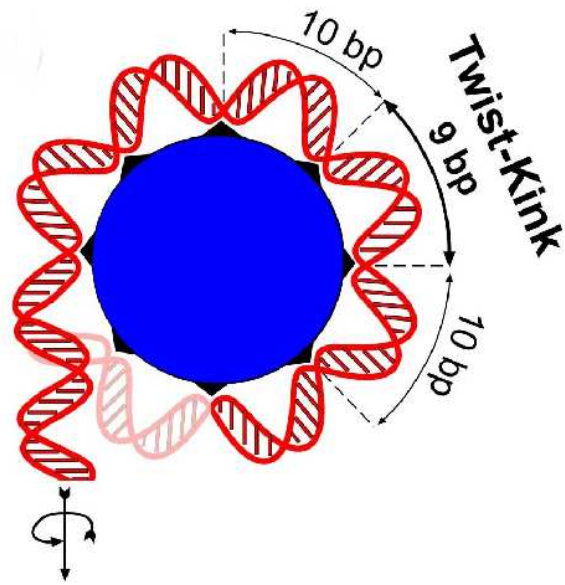
## Genetic information and biosynthesis:

- ▶ Chemical design of a protein (1 gen = 1 protein) is stored in the DNA which (for eukaryotes such as animals and plants) resides in the cell nucleus.
- ▶ DNA is subdivided into chromosomes (DNA double helixes wrapped around histones (chromatin proteins)) each of which carries quite a number of genes<sup>1</sup>.
- ▶ Chromosomes cannot leave the nucleus but the ribosomes where biosynthesis takes place are located outside the nucleus.

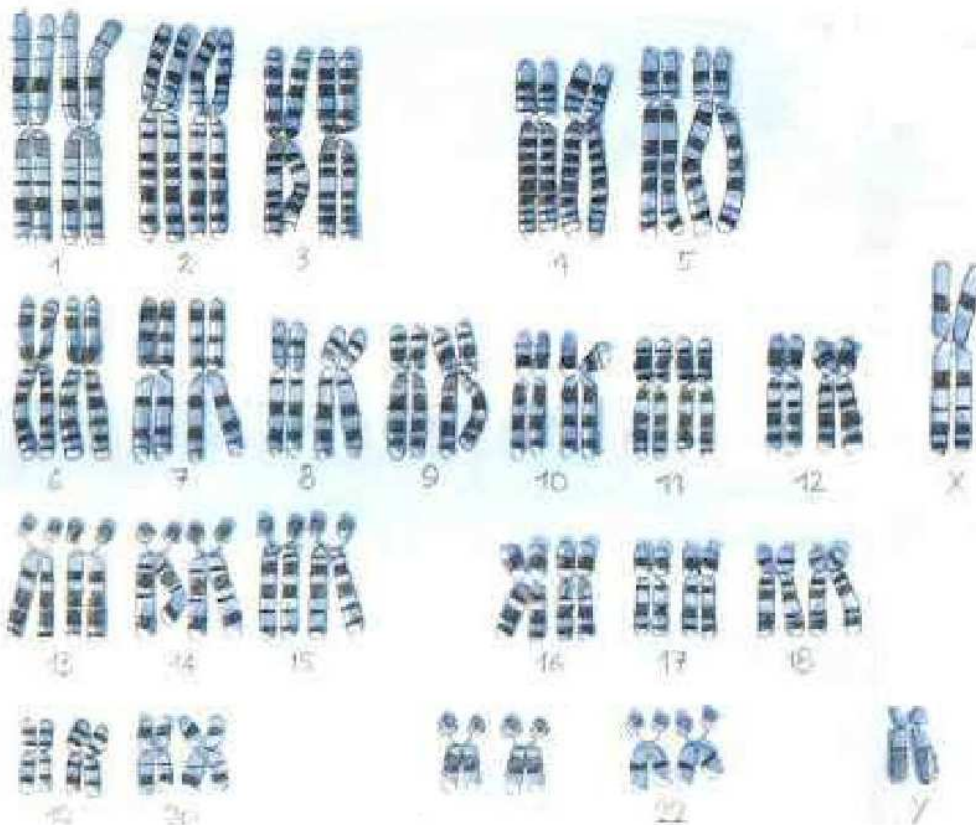
⇒ We need a means of transportation for genetic information.

---

<sup>1</sup>In most organisms the different (homologous) chromosomes occur in pairs where one is inherited from the mother and one from the father.





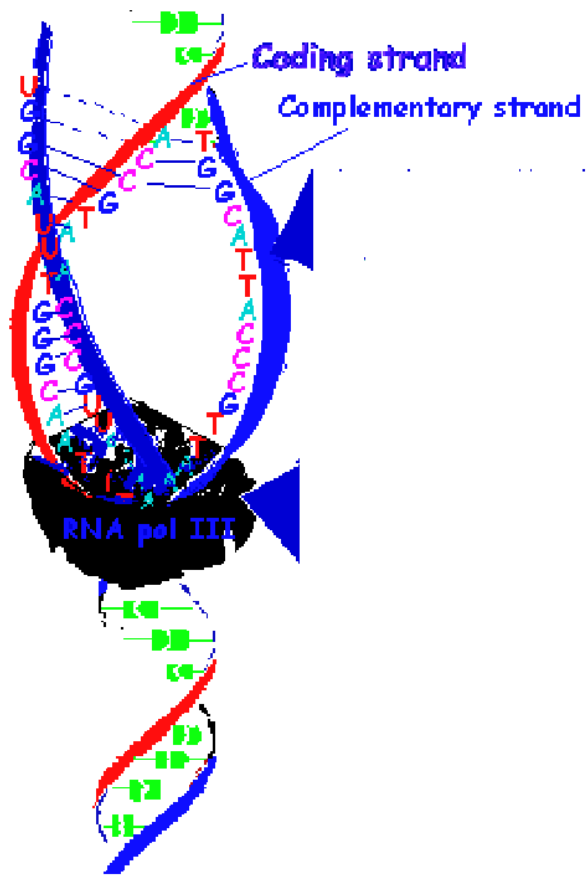


## Transcription:

(Task: build a copy of the genetic information)

- ▶ RNA polymerase (an enzyme) binds to a promoter in DNA and forces it to unwind and to produce a small open complex;
- ▶ synthesis begins on only the template strand and a (complementary) copy of the gene is produced;
- ▶ non-coding regions within the resulting RNA are removed (splicing);

⇒ Messenger RNA (mRNA for short), which leaves the nucleus and carries the information needed to the ribosomes.

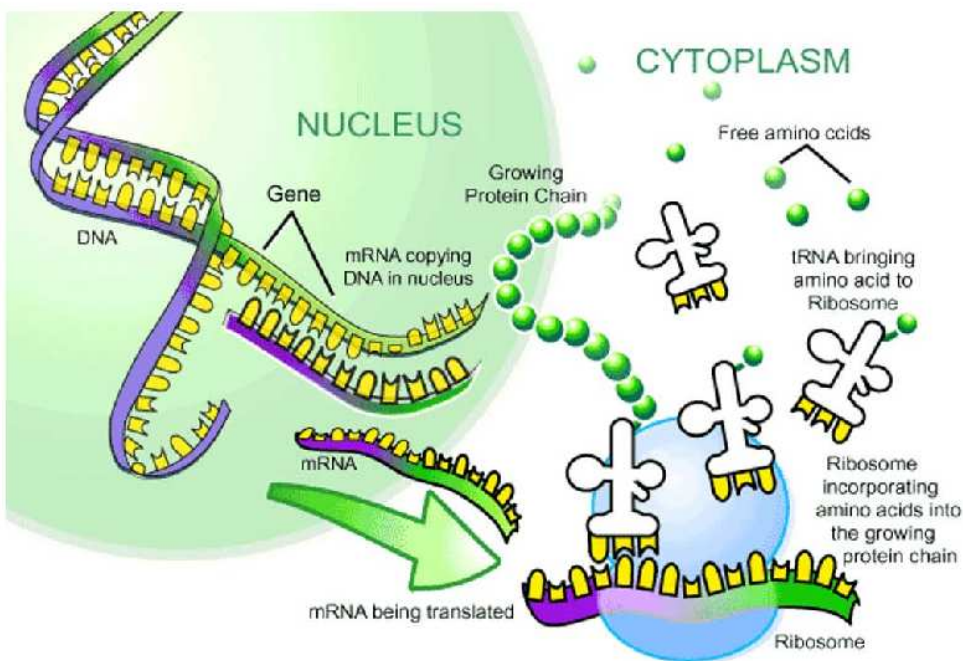


## Translation:

(Task: *translate* mRNA into sequence of amino acids)

- ▶ In each case, three successive bases (called a **codon**) are coding a single amino acid;
- ▶ the transfer RNA (**tRNA** for short) – another type of RNA – possesses a base triple complementary to the codon called **anticodon**;
- ▶ tRNA is folded onto itself and may bind an amino acid according to its anticodon;
- ▶ within the ribosomes the mRNA is processed codon after codon, while a tRNA-molecule always furnishes the next amino acid needed;
- ▶ this process ends, once a codon corresponding to the STOP-signal occurs.

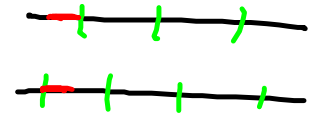
1. position	2. position				3. position
	G	A	C	U	
G	Gly	Glu	Ala	Val	G
	Gly	Glu	Ala	Val	A
	Gly	Asp	Ala	Val	C
	Gly	Asp	Ala	Val	U
A	Arg	Lys	Thr	Met	G
	Arg	Lys	Thr	Ile	A
	Ser	Asn	Thr	Ile	C
	Ser	Asn	Thr	Ile	U
C	Arg	Gln	Pro	Leu	G
	Arg	Gln	Pro	Leu	A
	Arg	His	Pro	Leu	C
	Arg	His	Pro	Leu	U
U	Trp	STOP	Ser	Leu	G
	STOP	STOP	Ser	Leu	A
	Cys	Tyr	Ser	Phe	C
	Cys	Tyr	Ser	Phe	U



# Lab techniques

## Basic methods:

- ▶ Heating: Divides a DNA-molecule in its two strands (denaturation as the reverse process to hybridation);
- ▶ Cutting the sugar phosphate backbone: sonic waves or vibrations are breaking the backbone at random positions, restriction enzymes are cutting predefined positions (every enzyme recognizes a specific sequence of nucleotides near which it cuts the DNA in blunt or sticky way).



## Amplifying (copying) DNA:

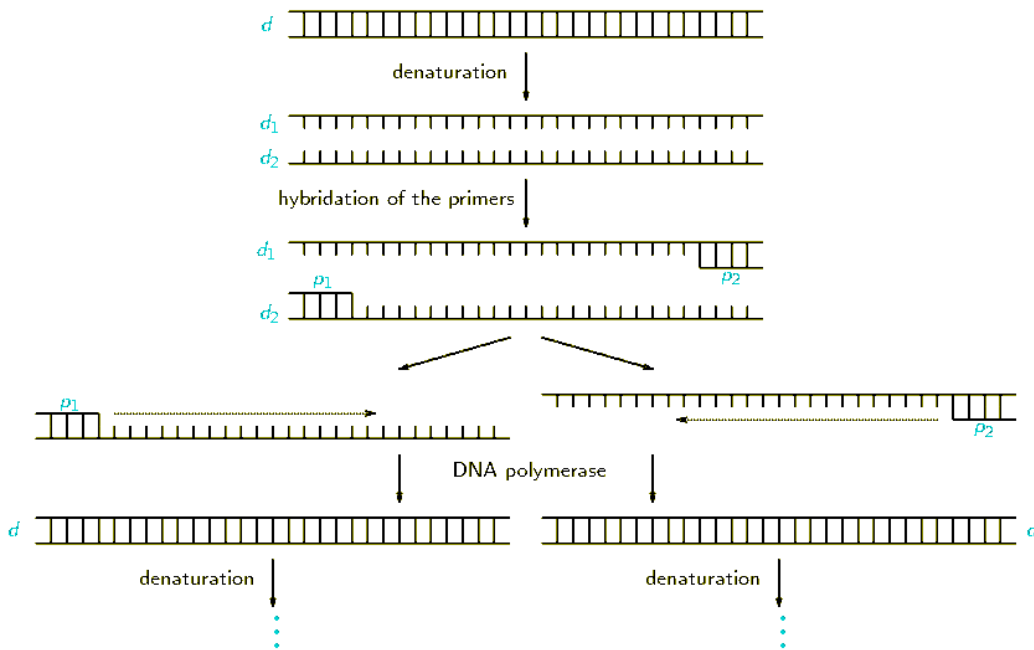
(polymerase chain reaction)

**Step 1:** Fill into a test-tube:

- ▶ The DNA  $d$  to be copied,
- ▶ the primer  $p_1$  and  $p_2$ , equal resp. complementary to the beginning resp. ending of  $d$ 's template strand,
- ▶ all nucleotides in sufficient magnitude, and
- ▶ DNA polymerase.

**Step 2:**

- ▶ Denaturate the DNA by heating.
- ▶ Let the DNA cool down. Along the way the primers hybridate at one end of each single strand.
- ▶ The DNA polymerase makes sure that both single strands are completed to two identical double-stranded copies of  $d$ .

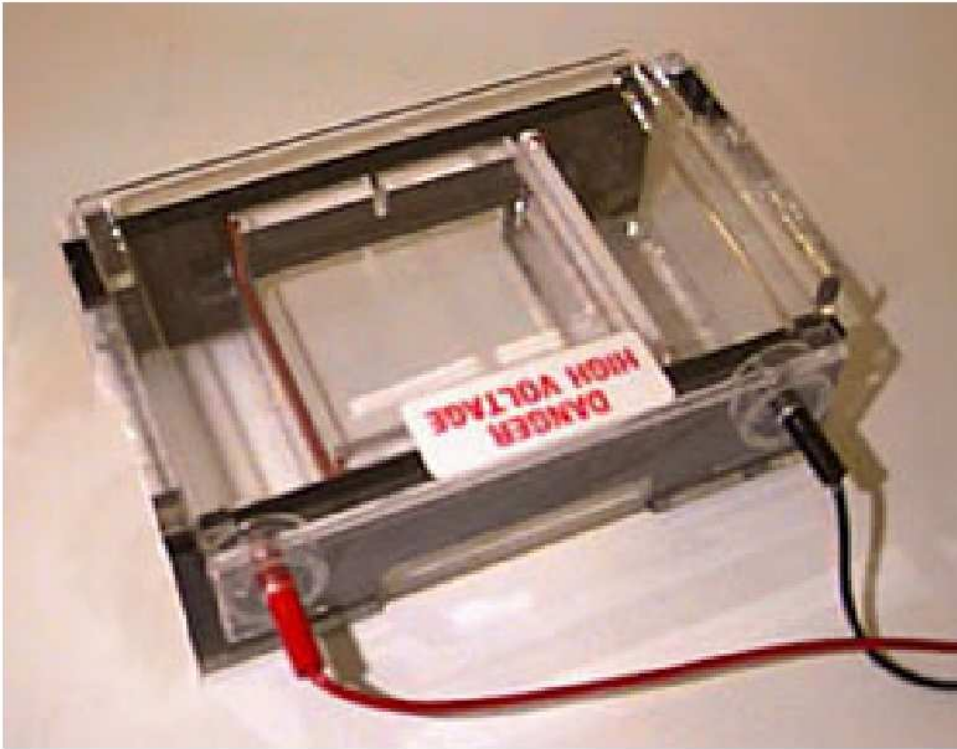


## Gel electrophoresis:

(Task: Sort DNA-molecules according to their length)

Background: DNA-molecules are negatively charged (due to their sugar-phosphate backbone) and therefore run towards the anode (positive pole) of an electric field.

- ▶ Place DNA samples into holes at one end of a *gel*;
- ▶ add an electrical current by which the DNA starts moving;
- ▶ short strands move through the holes in the gel more quickly than long strands (due to a smaller friction). Over time, the shorter strands therefore will move farther away from the starting point than the longer strands.
- ▶ by adding molecules of known sizes to a distinguished hole of the gel not only sorting is possible but we are able to determine the molecules sizes as well.



## Dideoxy sequencing (Sanger method):

(Determine the primary structure of  $d$ )

**Step 1:** Generate many copies of the DNA single strand  $d$ .

**Step 2:** Distribute these copies among four test-tubes  $A$ ,  $C$ ,  $G$  and  $T$ .

**Step 3:** Fill into test-tube  $X \in \{A, C, G, T\}$  all nucleotides beside the one with base  $X$ .

**Step 4:** Fill into test-tube  $X \in \{A, C, G, T\}$  in an appropriate ratio

- ▶ nucleotides, with base  $X$ , and
- ▶ nucleotides, with a chemically modified version of base  $X$ , at which the DNA polymerase cannot continue with completing the strand.

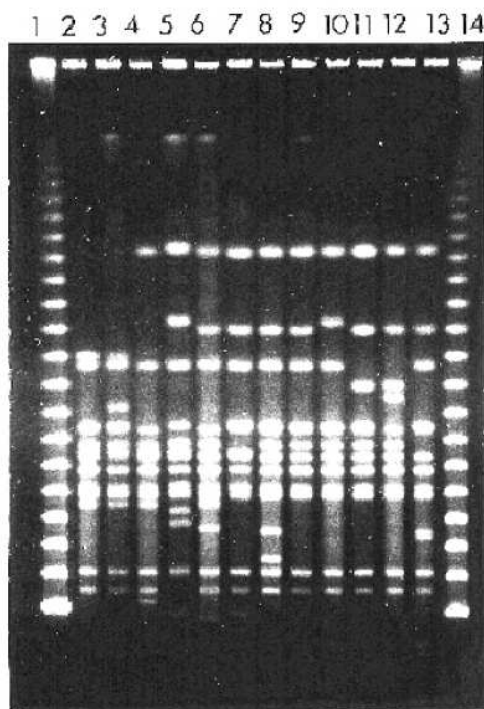
**Step 5:** Fill into all the test-tubes the primer for  $d$  and DNA polymerase.

Since test-tube  $X \in \{A, C, G, T\}$  contains the modified nucleotides with base  $X$  the tube will eventually contain all the incomplete double strands ending with  $X$  with high probability.

**Step 6:**

- ▶ Place the content of the four test-tubes into four different holes of a gel and run a gel electrophoresis.
- ▶ Since the molecules with different completion run on different speed through the gel, we can determine for each test-tube  $X$ , at which position hybridization has been canceled by means of the modified nucleotide with base  $X$ .
- ▶ If a molecule of test-tube  $X_1$  has moved furthest,  $d$  starts with the complement of  $X_1$ ; if a molecule of tube  $X_2$  went second furthest the second nucleotide of  $d$  is of type  $\overline{X_2}$  ...

⇒ We know which bases show up within  $d$  at all the different positions.





## Step 6:

- ▶ Place the content of the four test-tubes into four different holes of a gel and run a gel electrophoresis.
- ▶ Since the molecules with different completion run on different speed through the gel, we can determine for each test-tube  $X$ , at which position hybridation has been canceled by means of the modified nucleotide with base  $X$ .
- ▶ If a molecule of test-tube  $X_1$  has moved furthest,  $d$  starts with the complement of  $X_1$ ; if a molecule of tube  $X_2$  went second furthest the second nucleotide of  $d$  is of type  $\overline{X_2}$  ...

⇒ We know which bases show up within  $d$  at all the different positions.

**Step 5:** Fill into all the test-tubes the primer for  $d$  and DNA polymerase.

Since test-tube  $X \in \{A, C, G, T\}$  contains the modified nucleotides with base  $X$  the tube will eventually contain all the incomplete double strands ending with  $X$  with high probability.

## Dideoxy sequencing (Sanger method):

(Determine the primary structure of  $d$ )

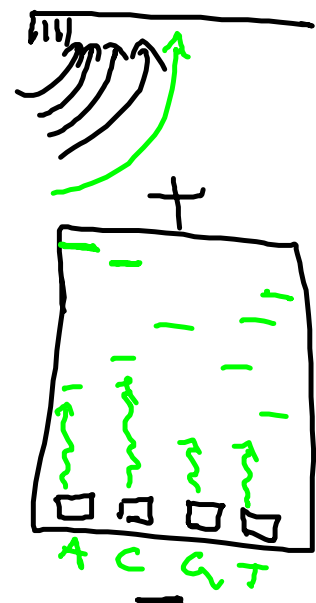
**Step 1:** Generate many copies of the DNA single strand  $d$ .

**Step 2:** Distribute these copies among four test-tubes  $A$ ,  $C$ ,  $G$  and  $T$ .

**Step 3:** Fill into test-tube  $X \in \{A, C, G, T\}$  all nucleotides beside the one with base  $X$ .

**Step 4:** Fill into test-tube  $X \in \{A, C, G, T\}$  in an appropriate ratio

- ▶ nucleotides, with base  $X$ , and
- ▶ nucleotides, with a chemically modified version of base  $X$ , at which the DNA polymerase cannot continue with completing the strand.



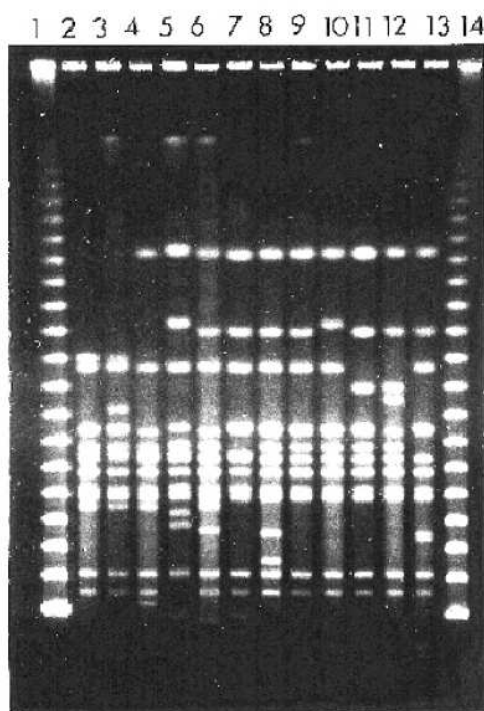
**Step 5:** Fill into all the test-tubes the primer for  $d$  and DNA polymerase.

Since test-tube  $X \in \{A, C, G, T\}$  contains the modified nucleotides with base  $X$  the tube will eventually contain all the incomplete double strands ending with  $X$  with high probability.

**Step 6:**

- ▶ Place the content of the four test-tubes into four different holes of a gel and run a gel electrophoresis.
- ▶ Since the molecules with different completion run on different speed through the gel, we can determine for each test-tube  $X$ , at which position hybridization has been canceled by means of the modified nucleotide with base  $X$ .
- ▶ If a molecule of test-tube  $X_1$  has moved furthest,  $d$  starts with the complement of  $X_1$ ; if a molecule of tube  $X_2$  went second furthest the second nucleotide of  $d$  is of type  $\overline{X_2}$  ...

⇒ We know which bases show up within  $d$  at all the different positions.

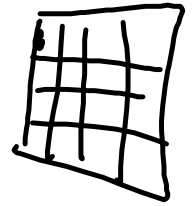


## DNA microarrays:

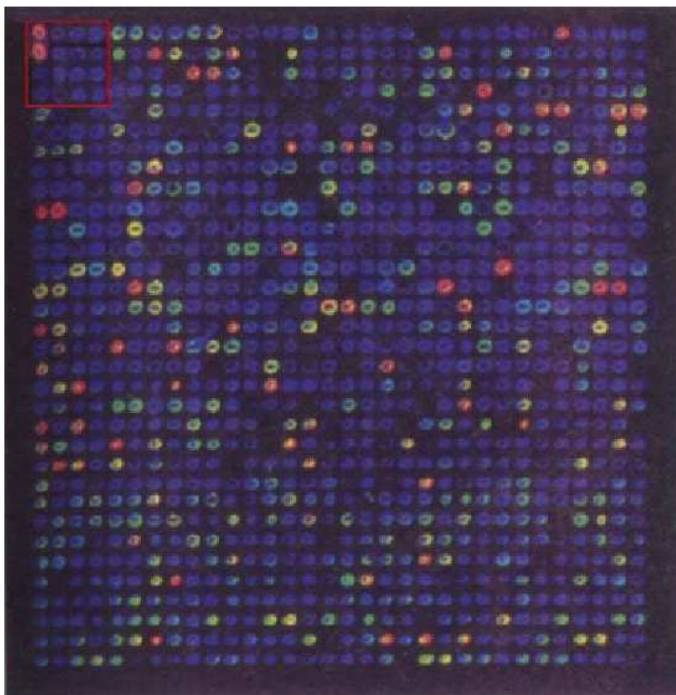
(Is  $t$  a subsequence of the unknown sequence  $s$ ?)

Idea: Perform a hybridization experiment for testing whether  $\bar{t}$  binds to  $s$  or not.

- ▶ For performing numerous such experiments at the same time, one uses so called **DNA micro arrays**, to which one can tie the complements  $\bar{t}_1, \dots, \bar{t}_n$  at distinguished positions.
- ▶ Afterwards, marked (fluorescent) copies of  $s$  are added; finally those not bonded to any of the  $t_i$  are washed away.
- ▶ The marking makes it possible to identify the subsequences showing up in  $s$ .



Attention: Experiments might be faulty (**false positives** and **false negatives** occur)!



# String Algorithms

## String-Matching:

### Definition

Given a text  $T \in \Sigma^*$  and a string  $P \in \Sigma^+$ , the string matching problem is to determine all  $s \in \mathbb{N}_0$ , satisfying:

$$(\exists v \in \Sigma^s, w \in \Sigma^*)(T = vPw).$$

The number  $s$  from this definition is named *shift*.

A shift is called *feasible*, if  $P$  is found at the respective place in  $T$ , otherwise  $s$  is called *infeasible*.

