Development of Complex Curricula for Molecular Bionics and Infobionics Programs within a consortial* framework**

Consortium leader

PETER PAZMANY CATHOLIC UNIVERSITY

Consortium members

SEMMELWEIS UNIVERSITY, DIALOG CAMPUS PUBLISHER

The Project has been realised with the support of the European Union and has been co-financed by the European Social Fund ***

**Molekuláris bionika és Infobionika Szakok tananyagának komplex fejlesztése konzorciumi keretben

***A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

INTRODUCTION TO BIOINFORMATICS

(BEVEZETÉS A BIOINFORMATIKÁBA)

CHAPTER 5

Sequencing and manipulating of DNA

(DNS szekvenálás és manipulálás)

Péter Gál

DNA sequencing

The development of the modern DNA sequencing techniques prompted the birth of bioinformatics. Due to the improving of the efficiency of the DNA sequencing techniques the size of the nucleotide databases has increased rapidly. DNA sequencing was and is the major source of biological data to be analysed by bioinformatics. DNA sequencing partially replaced protein sequencing, since it is much harder to sequence a protein than its gene or cDNA. The continuous improving of the sequencing techniques is a challenge for bioinformatics.

In 1972, a major achievement of DNA sequencing was to determine the sequence of a 174-bp-long DNA molecule.

This was featured on the front page of the prestigious scientific journal Nature. At that time the sequencing of the entire human genome (3.2×10^9 bp) would have taken more than one million years.

Now this task could be accomplished in a few weeks.

Sequencing the human genome (the Human Genome Project) was one of the greatest scientific achievement of the 20th century.

Till now the genome of at least 6000 different species has been sequenced and the data have been put to the data bases.

Landmarks in the Human Genome Project

1953 Watson and Crick published on the structure of the double stranded DNA molecule (Nature, April 25th)

„*We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.*”

„*It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.*”

The most famous understatements of the 20th century.

1975 The advent of the modern DNA sequencing techniques

At that time two methods of DNA sequencing were developed independently.

Both methods generates DNA fragments that represent the entire sequence. The fragments are resolved by polyacrylamide gel electrophoresis under denaturing conditions.

The Maxam-Gilbert method uses chemical modifications of the bases while the Sanger method uses DNA polymerase enzyme to copy fragments form the template DNA strand.

The Sanger method was for a long time the leading method for DNA sequencing.

1977 The bacteriophage ΦX-174 was sequenced.

This 5.4 kbp DNA represents the first complete genome ever sequenced.

1981 The DNA of the human mitochondrion was sequenced.

It consist of 16569 bp.

1984 The first sequence of a human virus, the Epstein-Barr virus (one of the eight known human herpes viruses), was determined.

The virus genome is 172281-bp-long.

1990 The International Human Genome Project was launched.

At the beginning this international project planned to sequence the human genome in 15 years.

1991 Craig Venter identifies active genes in the genome by sequencing the initial portions of complementary DNA (cDNA). cDNA is made from mRNA by means of reverse transcription. Every short piece of cDNA can be considered as a tag of the whole gene. Hence the name Expressed Sequence Tags (ESTs).

1992 The complete low resolution linkage map of the human genome was determined.

1992 The *Caenorhabditis elegans* (nematode worm) sequencing project was launched.

This year two new research center were established.

In the framework of the Human Genome Project the Sanger Center for large-scale genomic sequencing was established in Hinxton, UK.

Craig Venter established The Institute for Genome Research (TIGR) for the commercial application of the sequence information derived form the genome sequencing projects. The major aim was identify genes that encodes potential drug targets for the pharmaceutical industry.

1995 The first bacterial genome was sequenced by TIGR.

The *Haemophilus influenzae* genome, the first genome of a free-living organism (1.8 million base pairs), was sequenced.

The same year the genome of *Mycoplasma genitalium*, the smallest genome of a self-replicating organism (582970 bp) was also sequenced by Craig Venter's company.

1996 The high-resolution map of human genome was established. The resolution of the map is approximately 600 000 bp.

The same year the first complete eukaryotic genome, the genome of the yeast *Saccharomyces cerevisiae*, was sequenced.

1998 Craig Venter's company, Celera, announced that they would finish the sequencing of the human genome by 2001. In response to that the sponsors of the Human Genome Project increased the funding of the Sanger Center.

The same year the complete genome sequence of the nematode worm *Caenorhabditis elegans* was published.

1999 At the end of this year Celera Genomics announced that they determined the genome sequence of Drosophila melanogaster.

The sequence data were released in spring 2000.

1999 The Human Genome Project announced its plan to finish the working draft of human genome (90% of genes sequenced to >95% accuracy) by 2001.

December 1, 1999 Sequence of first complete human chromosome published.

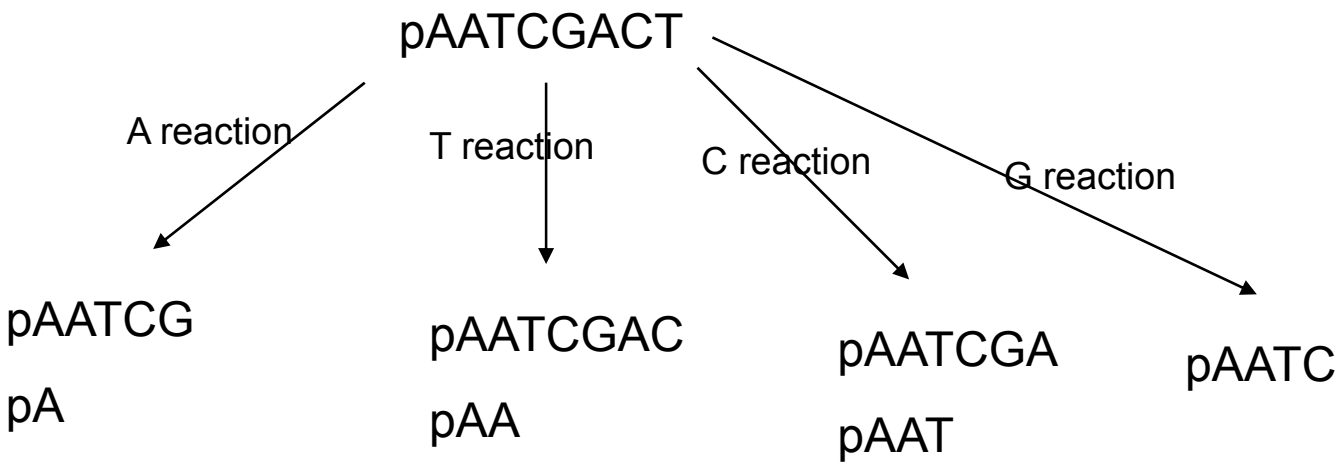
June 26, 2000 Joint announcement (Prime Minister of the United Kingdom, Tony Blair, and President of the United States, Bill Clinton) of the completion of the draft of the Human Genome. Scientists form the Human Genome Project and from Celera Genomics were also present.

2003 Completion of high-quality human genome sequence by the public consortium.

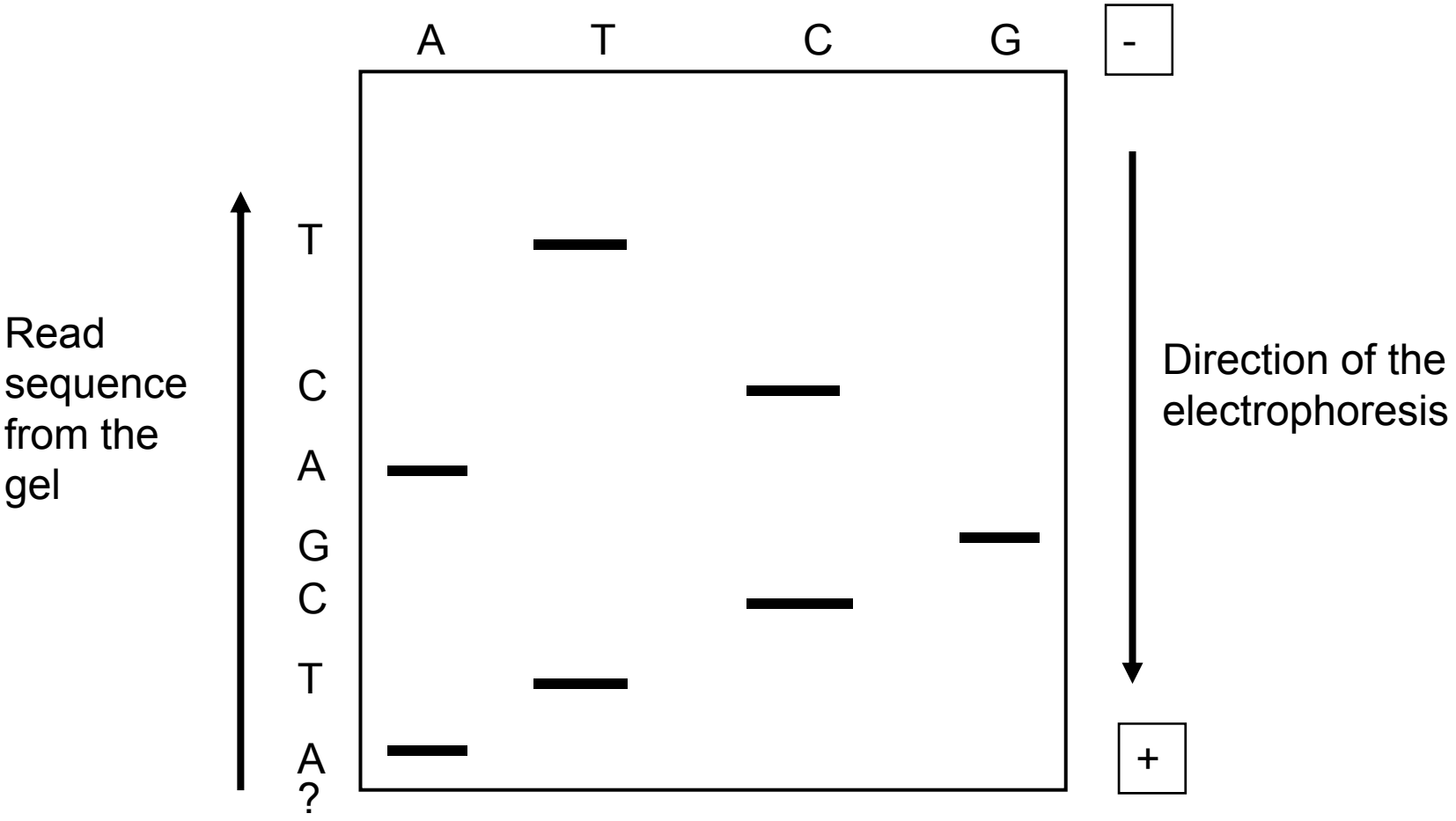
DNA sequencing methods

Maxam-Gilbert method

Principle: To generate four sets of labeled fragments from the DNA to be sequenced by chemical reactions.



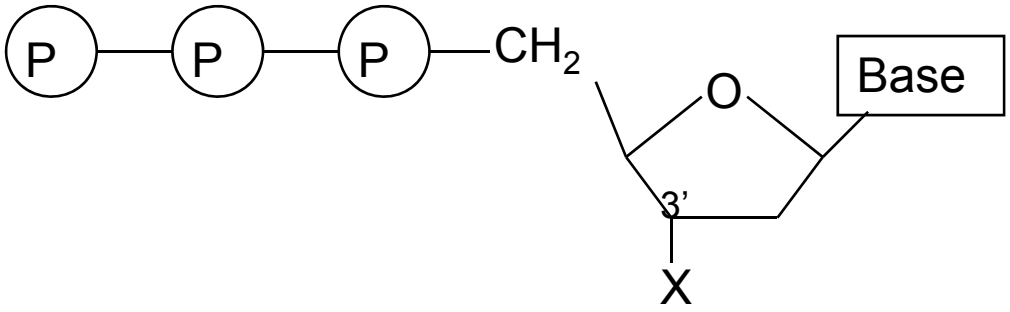
Separation of the labaled fragment by gel electrophoresis



The Maxam-Gilbert sequencing was the method of choice for sequencing shorter DNA molecules, especially chemically synthesized oligonucleotides. The main drawbacks of the method are the need of purified homogeneous DNA and the usage of toxic chemicals.

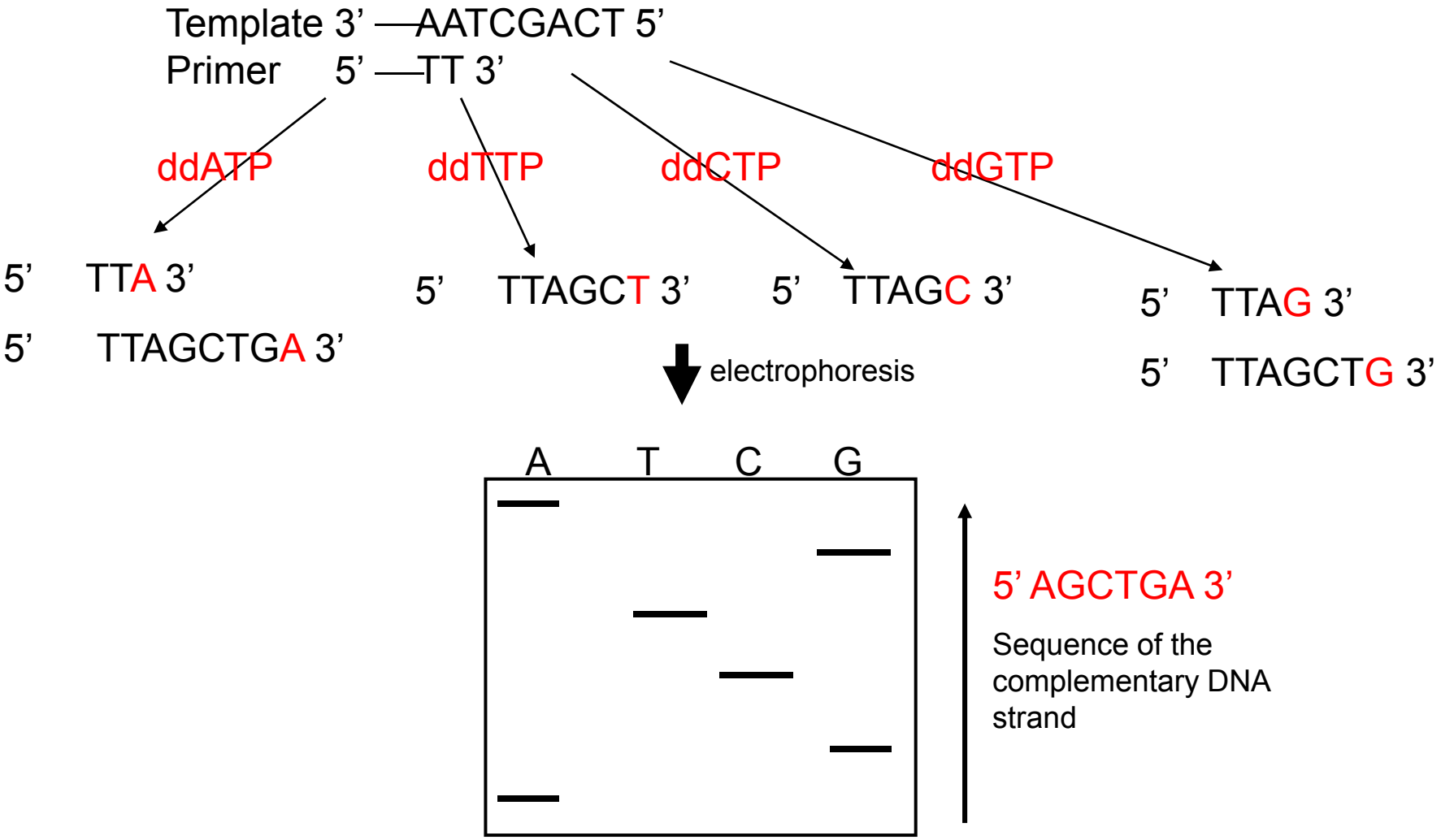
The Sanger method uses basically the same principle as the Maxam-Gilbert (i.e. generation of base specific fragments), however this method generates the base specific DNA fragments by enzymatic synthesis of the complementary DNA strand. To stop the synthesis of the DNA strand at a specific point it uses dideoxynucleoside triphosphates (ddNTP).

Dideoxynucleoside triphosphates (ddNTPs) are deoxynucleoside triphosphate (dNTP) analogues where there is no OH group at the 3' position of the ribose.



X= OH→ dNTP
X= H→ ddNTP

The Sanger method of DNA sequencing uses DNA polymerase enzyme for synthesizing the complementary DNA strand. DNA polymerases cannot initiate the synthesis of a new DNA strand alone. They need a primer oligonucleotide, that hybridize to the single-stranded DNA template and the DNA polymerase adds the next incorporating nucleotide to the 3' OH of the new DNA strand. In case of the incorporation of a dideoxy nucleotide, there will be no 3' OH, consequently the synthesis of the new DNA strand terminates. If we use all the four ddNTP in separated reactions (a certain ratio of dNTP and ddNTP) we can generate all possible DNA fragments.

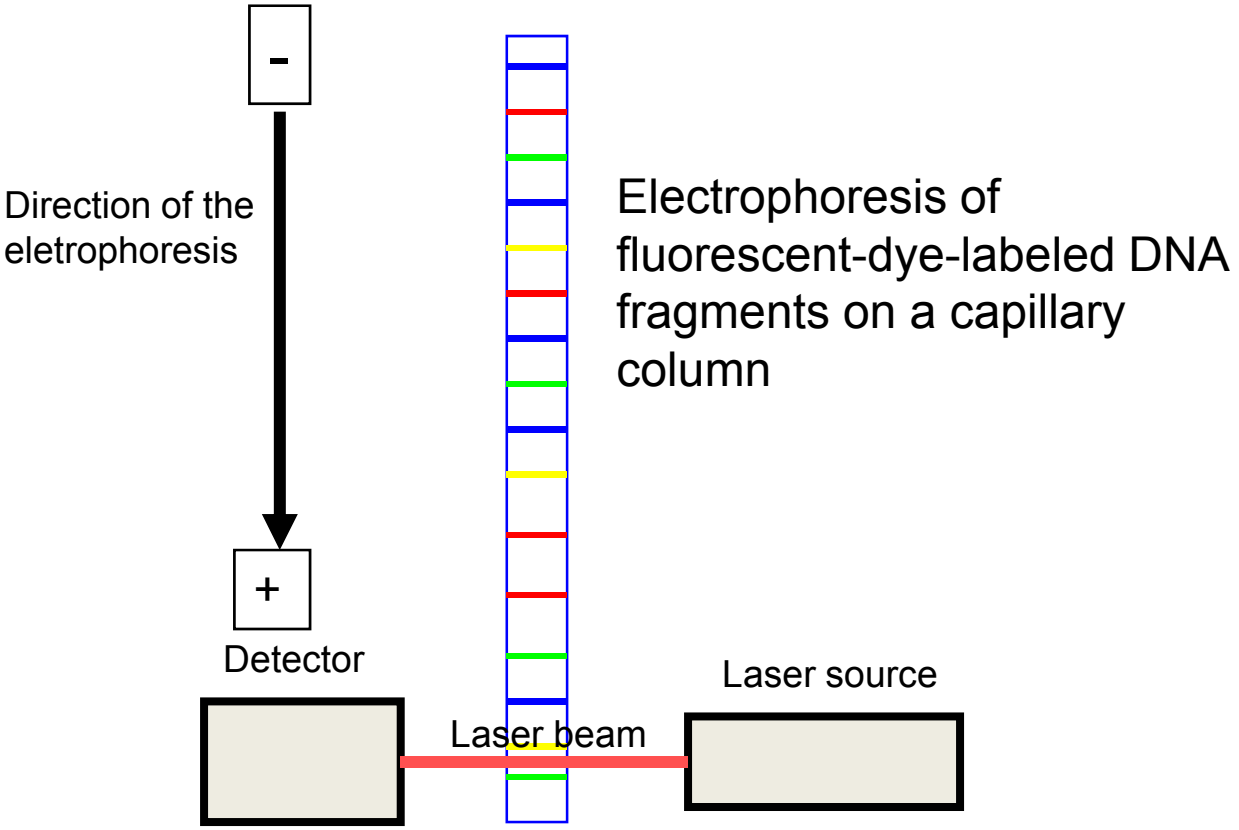

Template 3’ —AATCGACT 5’
Primer 5’ —TT 3’
ddATP
ddTTP
ddCTP
ddGTP
5’ TTA 3’
5’ TTAGCTGA 3’
5’ TTAGCT 3’
5’ TTAGC 3’
5’ TTAG 3’
5’ TTAGCTG 3’
electrophoresis
A T C G
5’ AGCTGA 3’
Sequence of the complementary DNA strand

The Sanger method become the most popular method of DNA sequencing, since it works with partially purified double stranded DNA (e.g. recombinant plasmid DNA) and it can be readily automated. The automated Sanger sequencing method was the workhorse of the Human Genome Project.

Now days the next generation sequencing (NGS) methods gradually replace the Sanger method.

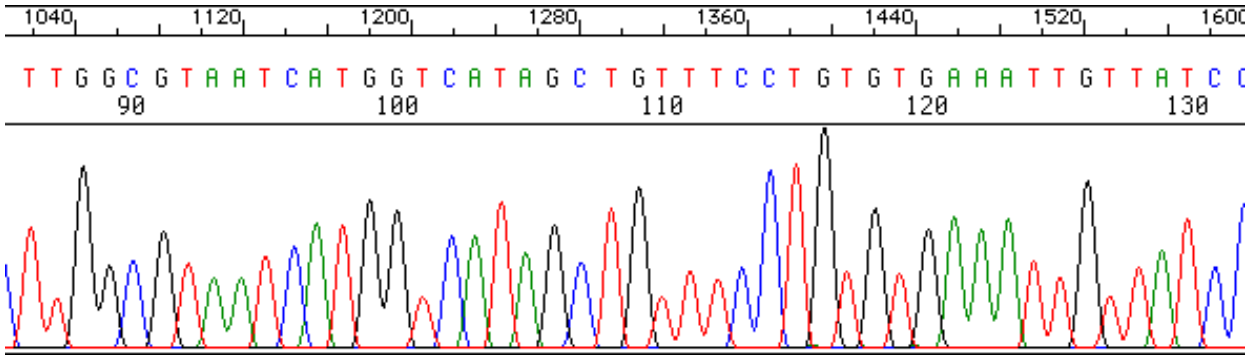
Frederick Sanger has been awarded two Nobel Prizes in Chemistry, one for protein sequencing (1958) and one for DNA sequencing (1980). (There are only four persons in history having two Nobel Prizes.)

Automation of the Sanger DNA sequencing



Electrophoresis of fluorescent-dye-labeled DNA fragments on a capillary column

Computer-generated result (chromatogram) of the automated Sanger DNA sequencing



Next generation sequencing

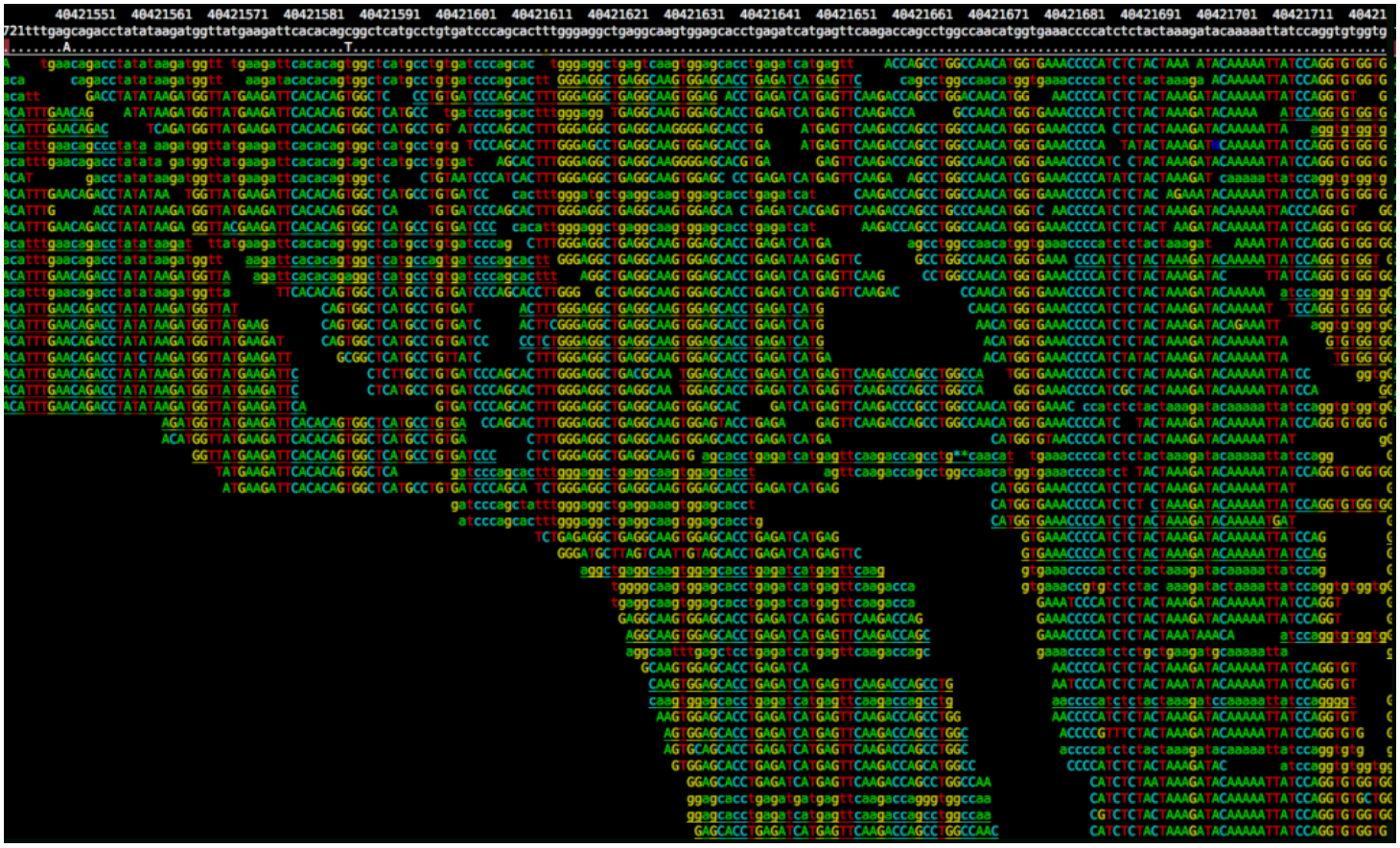
Recently new methods of DNA sequencing appeared that are an order of magnitude faster than the Sanger's one.

Moreover they are cheaper and have deep enough coverage to look at genomes of individuals or even tissues.

It makes possible to read the sequence of several hundred millions DNA molecules at the same time.

The rise of the next generation sequencing is a challenge for bioinformatics. The first task is to assemble a whole sequence (e.g. a chromosome) from tens of millions of short sequencing fragments.

A tipical output of an NGS experiment



Steps of a high-throughput sequencing experiment

1.) Sample preparation: the large DNA molecule is fragmented to small single-stranded pieces.

2.) The individual DNA molecules are immobilized on small beads.

3.) The synthesis of the complementer strand takes place on the bead. Every incorporation of a nucleotide gives a flash of fluorescent light which is characteristic to the given nucleotide.

4.) The process is monitored by an extremely sensitive digital camera. The camera with the data processing device is capable of monitoring millions of bead, that is the synthesis of millions individual DNA molecules.

The primary data are the pictures with the millions of fluorescent flashes which can be followed real-time online.

The colors of the flashes are then translated into DNA sequences (secondary data) that correspond to the complementary strand of the DNA immobilized on the bead.

Finally, the sequence of the original full-length DNA molecule should be assembled from the tens of millions of short sequencing fragments.

It took ten years and approximately 300 million dollars for the Human Genome Project to sequence the entire human genome in the last century. Now an individual's genome can be sequenced within a few weeks for several thousand dollars. The aim is to reduce the price under a thousand.

Strategies for genome sequencing

1.) Systematic sequencing:

Mapping the DNA, sequencing relatively long pieces of DNA, and then assemble the entire DNA molecule (chromosome) by using the map.

It is a slow but reliable method. It is labor-intensive.

2.) Shotgun sequencing

Fragmentation of the large DNA molecule into small pieces, sequencing the fragments, assemble the whole sequence by using the overlapping sequence.

Problems with the shotgun sequencing:

We have to read at least 10 times more nucleotides than the length of the original DNA in order to achieve a gap-free sequence.

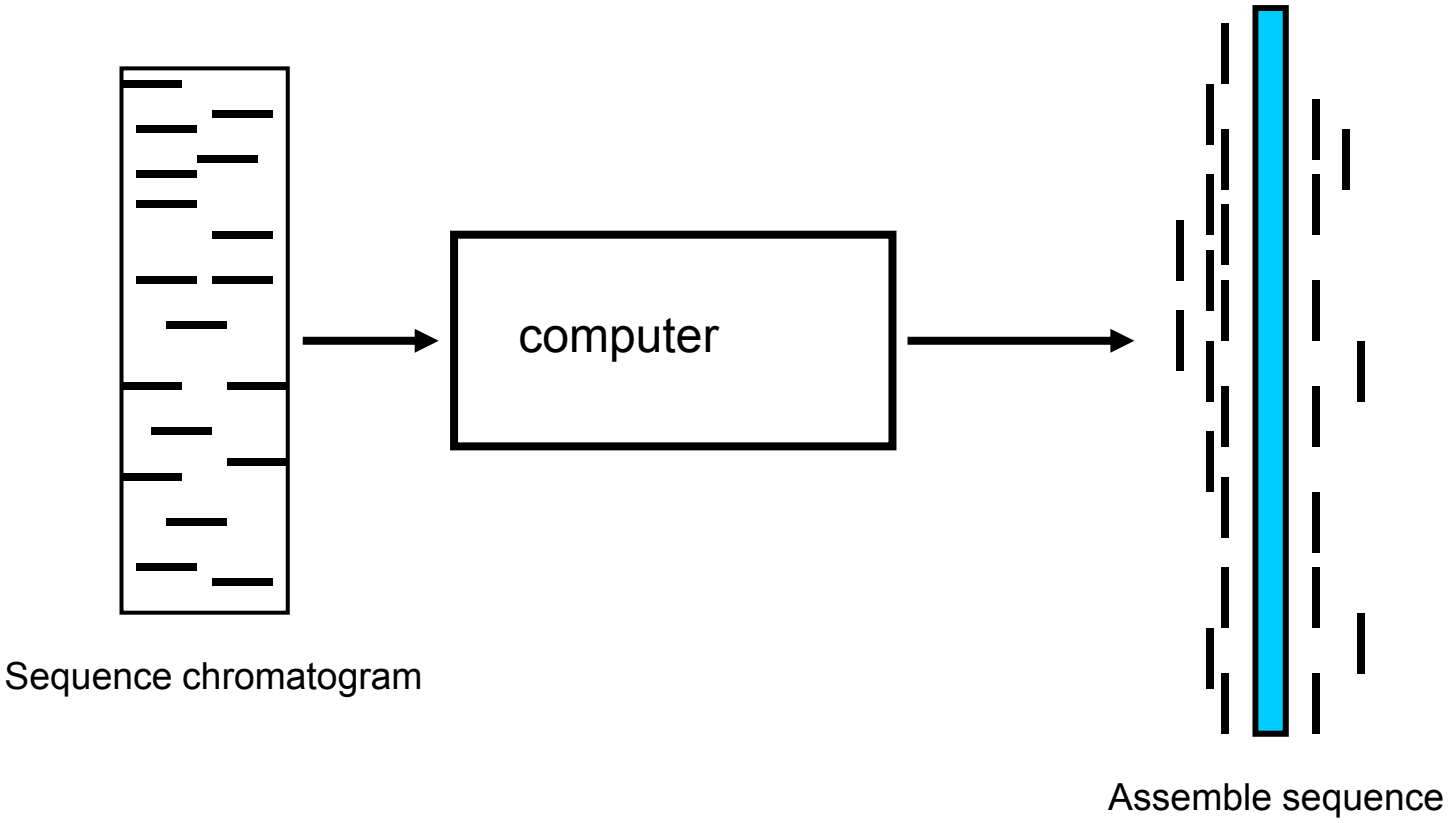
The repetitive sequences make the assembly ambiguous. (We need a map anyway.)

We need a lot of computer-power: (memory and running time).

It works perfectly with small (prokaryotic) genomes.

In case of eukaryotic genome the high ratio of repetitive sequences causes problems.

Procedure of the shotgun sequencing



DNA cloning

Many DNA sequence data that can be found in the databases are sequences of cloned DNA molecules. Usually cloning and/or amplification of a DNA (or RNA) molecule precludes the sequencing. Cloning is the prerequisite of obtaining certain type of sequence information, such as cDNA sequences, EST libraries.

Alternative names of DNA cloning: recombinant DNA technology; genetic engineering.

DNA cloning is basically manipulation of the DNA by special enzymes.

The most important enzymes are the restriction endonucleases.

These enzymes are of bacterial origin and recognize 4-6 nucleotide-long sequences within the double stranded DNA and cut the DNA chain.

Several hundred restriction endonucleases are known.

We can create the restriction map of the DNA molecule which is also called the physical map of DNA.

The products of the restriction digestion (restriction fragments) can be inserted into suitable carrier DNA molecules (vectors) and then we can seal the DNA chains by DNA ligase enzyme.

The resulting artificial DNA molecule is called recombinant DNA.

The recombinant DNA can be introduced into suitable host cells (e.g. Escherichia coli), which can be amplified up to millions of copies. The bacterial colony that contains the copies of a distinct recombinant DNA molecule is called a clone. The term „clone" usually means the recombinant DNA itself.

The recombinant „clone" contains enough DNA for any standard analysis (e.g. sequencing).

Restriction endonucleases

AluI produces blunt ends

EcoRI and HindIII produce „sticky" ends

Recombinant DNA

Molecule A

5'…GGATCC…3'

3'…CCTAGG…5'

Molecule B

5'…GGATCC…3'

3'…CCTAGG…5'

↓ Digest with the same restriction endonuclease, BamHI

5'…G

3'…CCTAG

GATCC…3'

G…5'

↓ Mix, seal with DNA ligase

5'…GGATCC…3'
3'…CCTAGG…5'

Recombinant DNA

Complementary DNA (cDNA)

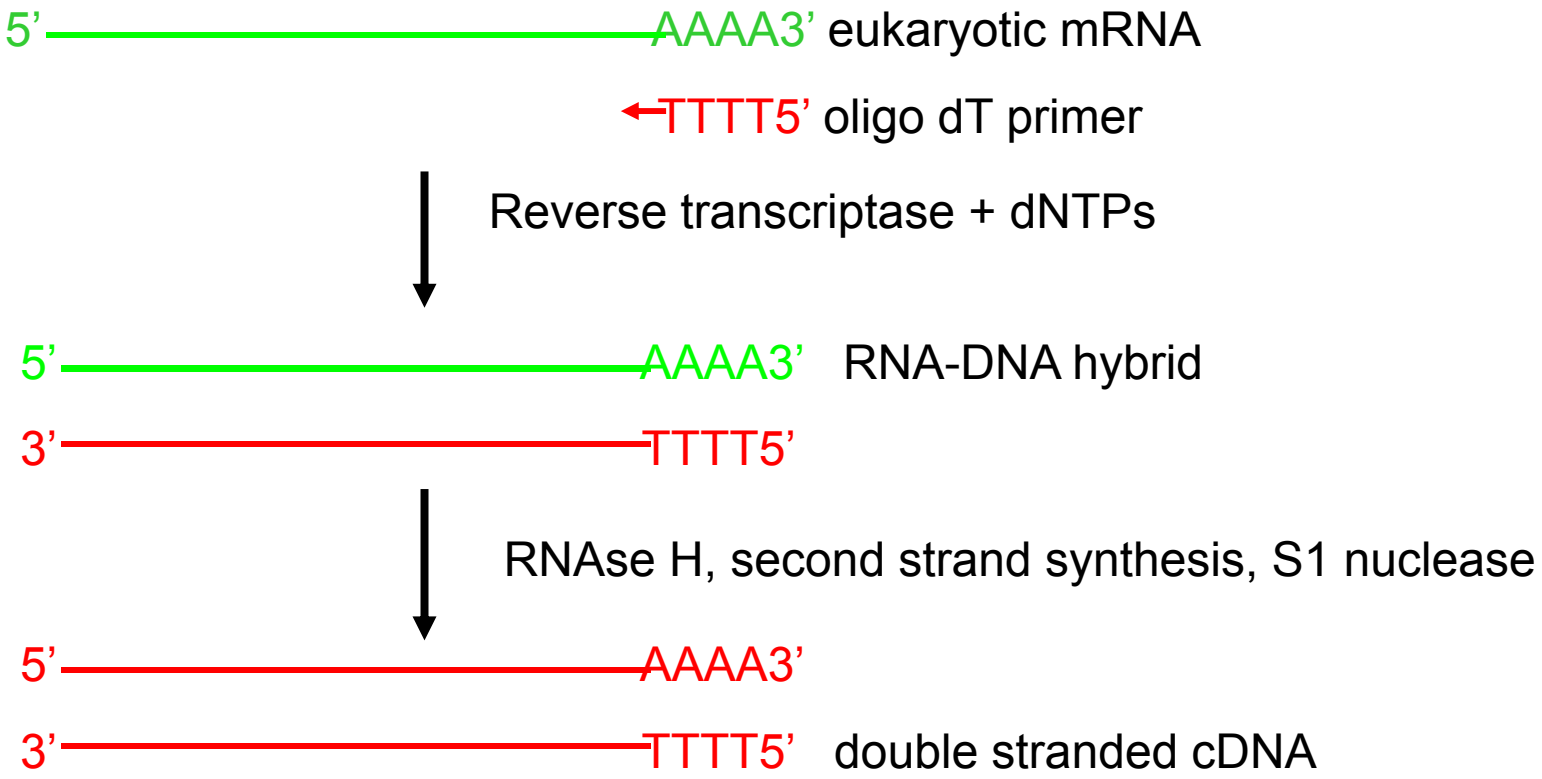
Since the eukaryotic mRNAs do not contain the introns, there is a direct relationship between their nucleotide sequence and the amino acid sequence of the proteins they encode.

RNA molecules, however are chemically unstable, they are not suitable for cloning purposes.

Solution: Making a DNA copy from the mRNA molecule. Reverse transcriptase enzymes can use RNA templates to synthetise DNA. The resulting DNA molecule is called complementary (copy) DNA or simply cDNA.

cDNA is suitable for cloning purposes and can be used for recombinant production of eukaryotic proteins in prokaryotic hosts (bacteria).

Synthesis of cDNA



The most frequently used host system in the recombinant DNA technology is the K12 strain of *Escherichia coli*.

It is a Gram-negative bacterium, that is well characterized at molecular level.

It is suitable for propagating recombinant DNA molecules.

The most frequently used vectors that can be used for introduction of foreign DNA into the *E. coli* cells are the plasmids and the bacteriophages.

Eukaryotic cells (yeast, insect cells, mammalian cells, etc.) are also used for expression of recombinant proteins but *E. coli* is always the system of choice for manipulation of DNA.

Vectors

Plasmids: Extra-chromosomal genetic elements in bacteria.

The size of a plasmid is typically 1-200 kbp.

They are circular, double stranded DNA molecules.

They can replicate independently form the bacterial chromosome.

The copy number of a plasmid inside the bacterial cell is determined by the origin of replication. It can range from a few copies up to several hundreds of molecules.

The modern plasmid vectors contain many artificial DNA segments for facilitating the cloning procedure (e.g. polycloning site, antibiotic resistance, single-stranded DNA replication origo. LacZ gene for blue-white selection, etc.).

Vectors

Bacteriophages: the viruses of the bacteria

They can replicate independently inside the bacterial cell and they tolerate the presence of foreign DNA inserts.

Advantages over the plasmids: They are their own efficient way to enter into the cell. The transformation process is very efficient.

The foreign DNA insert can be much bigger than in the case of bacteria (8-24 kbp).

Disadvantage: The laboratory process is more difficult and expensive.

The most frequently used bacteriophage vectors are the λ-phage and the M13 phage.

Vectors

M13 phage is the vector of special applications, such as preparing single-stranded DNA and displacing recombinant proteins on the surface of the virus for in vitro selection (phage-display).

Other vectors:

Cosmids, BACs (bacterial artificial chromosomes), YACs (yeast artificial chromosomes).

They are suitable for cloning of large pieces of DNA:

Kozmid	30-45 kbp
BAC (bacterial artificial chromosome)	120-300 kbp
YAC (yeast artificial chromosome)	250-400 kbp

In order to identify the cloned DNA we can use several methods including colony hybridization, restriction mapping, direct sequencing and polymerase chain reaction (PCR).

PCR is the cell-free method of DNA amplification.

We can clone DNA without using vector and host cells.

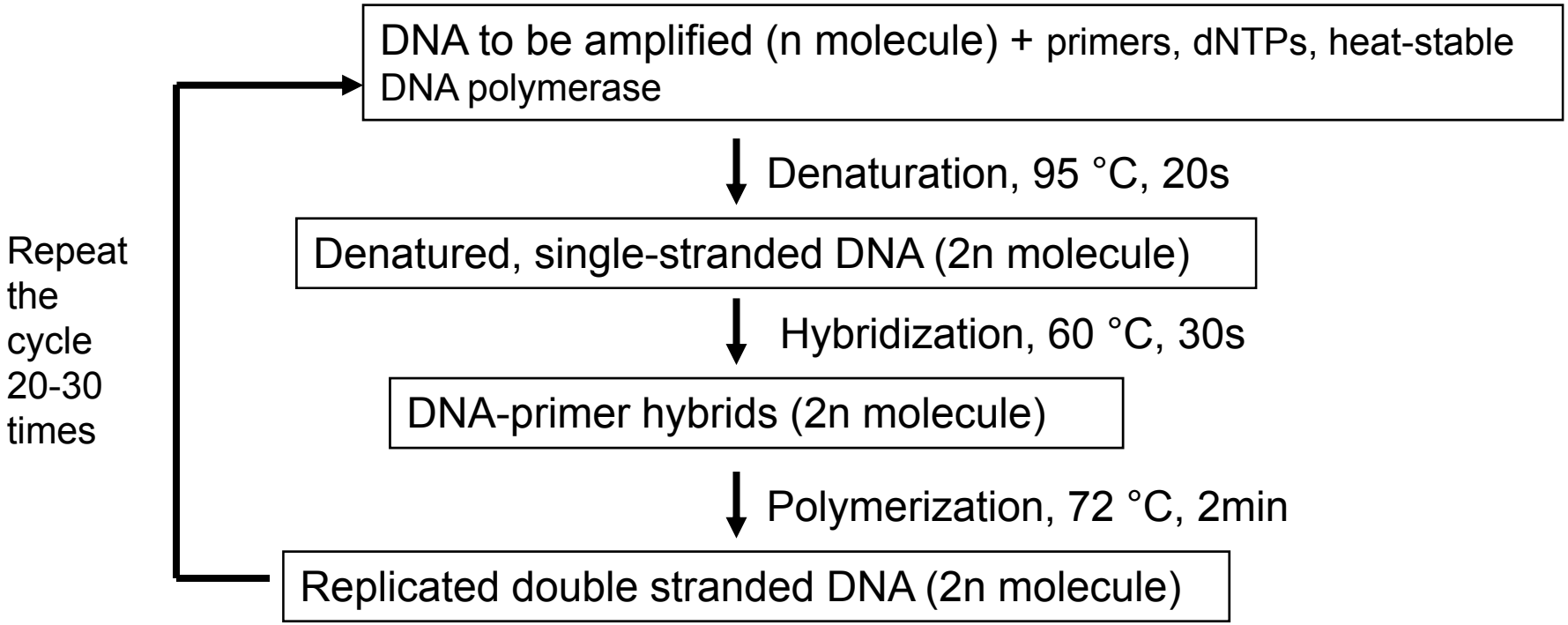
PCR resembles to the *in vivo* DNA replication.

Any DNA sequence can be amplified.

The only restriction is that we should know the short flanking sequences around the target DNA.

At present PCR is the most frequently used method for DNA cloning.

Polymerase chain reaction



The power of the PCR

Cycles	Copies of DNA molecules
1	2
2	4
4	16
10	1024
15	32768
20	1048576
25	33554432
30	1073741824

PCR

Primer design: the most critical step at PCR.

Critical parameters are: length, melting point, cross hybridization, secondary structural elements in the DNA

At present many softwares are available for PCR primer design.

One real limitation: we should know the sequence flanking the DNA to be amplified.

This limitation however can be overcome by certain cases: e.g. we can design primers for the vector to amplify the cloned DNA fragment, we can use oligo dT primers at the 3' end of the cDNA, we can synthetise homopolymer region on the 3' end of the DNA, etc.

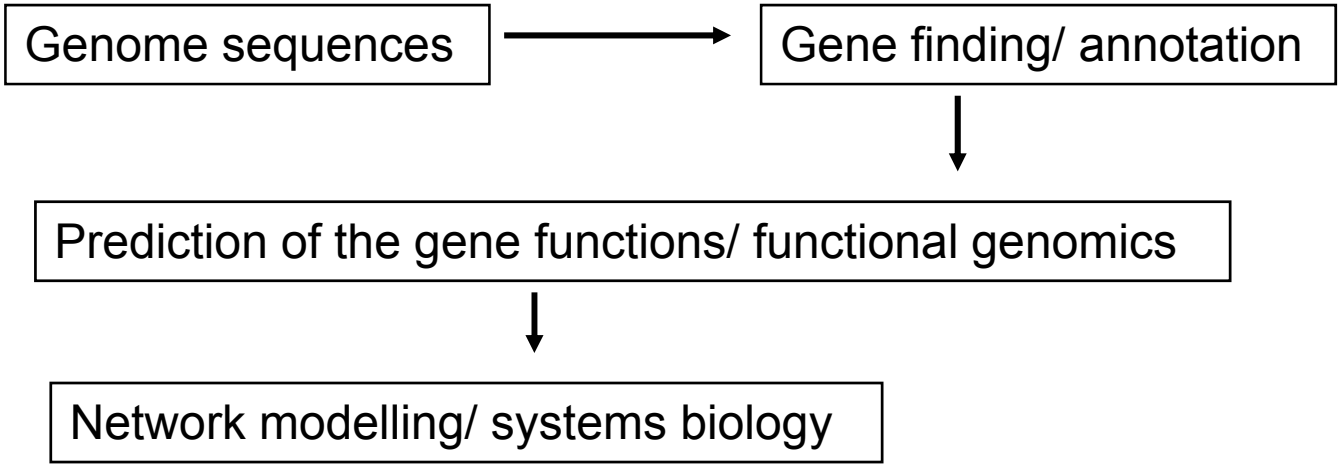
Cloning in the genomic age

The first step is to find the sequence of the gene to be cloned in a database (e.g. gene bank).

Knowing the sequence we have several choices:

- Many genes can be purchased from biotech companies ligated into a vector.
- We can design primers to amplify the DNA. In this case we need suitable template DNA.
- We can synthesize the gene. The advantages of the artificial genes are that we can design the restriction map of the DNA and we can tune the codon usage for recombinant protein expression.

Due to the recombinant DNA technology and the rapid develpoment of the modern sequencing methods the size of the (nucleotid) sequence databases are expanding very rapidly.



Problem: we have a plethora of predicted potential gene sequences from the recently sequenced genomes, however we do not know for sure which one is really encoding protein, let alone the structure and the function of the encoded proteins.

Even in the case of a „simple" organism there are many genes with unidentified function (orphan genes).

Example: The genome of the yeast (*Saccharomyces cerevisiae*) contains more than 6000 genes. Before genome sequencing approximately 2000 genes have been characterized experimentally. Another 2000 genes was not know before, however their functions can be predicted by homology. The remaining 2000 genes however do not show any homology with other known genes (orphan genes) therefore their function is elusive.