DNA SEQUENCING
CHEMICAL METHOD AND ENZYMATIC METHOD

DNA SEQUENCING

- The process of determining the order of bases adenine (A), thymine (T), cytosine (C), and guanine (G) along a DNA strand.
- All the information required for the growth and development of an organism is encoded in the DNA of its genome.
- So, DNA sequencing is fundamental to genome analysis and understanding the biological processes in general.

TECHNICAL BREAKTHROUGH FOR DNA SEQUENCING

- In 1977, two separate methods for the large-scale sequencing of DNA were Included:
  - Chemical cleavage method by Maxam and Gilbert
  - Enzymatic chain termination method by Sanger

- Of these two methods, Sanger method is more popular. Without changing the underlying concept of both methods, some improvements have been done over the years by applying different strategies, by developing various modifications and by automation.
- As a result, a very large scale sequencing has become feasible, e.g. *E. coli*, *Saccharomyces cerevisiae*, Human Genome Project etc.

CHEMICAL CLEAVAGE METHOD

- This method uses double-stranded DNA samples.
- Involves modification of the bases in DNA followed by chemical base-specific cleavage.
- Sequences DNA fragments containing up to ~500 nucleotides in length.

STAGES:

1. The double-stranded fragment to be sequenced is isolated and radioactively labeled at the 5'-ends with ^{32}P.

2. The fragment is then cut with restriction enzyme and thus the label is removed from one end.

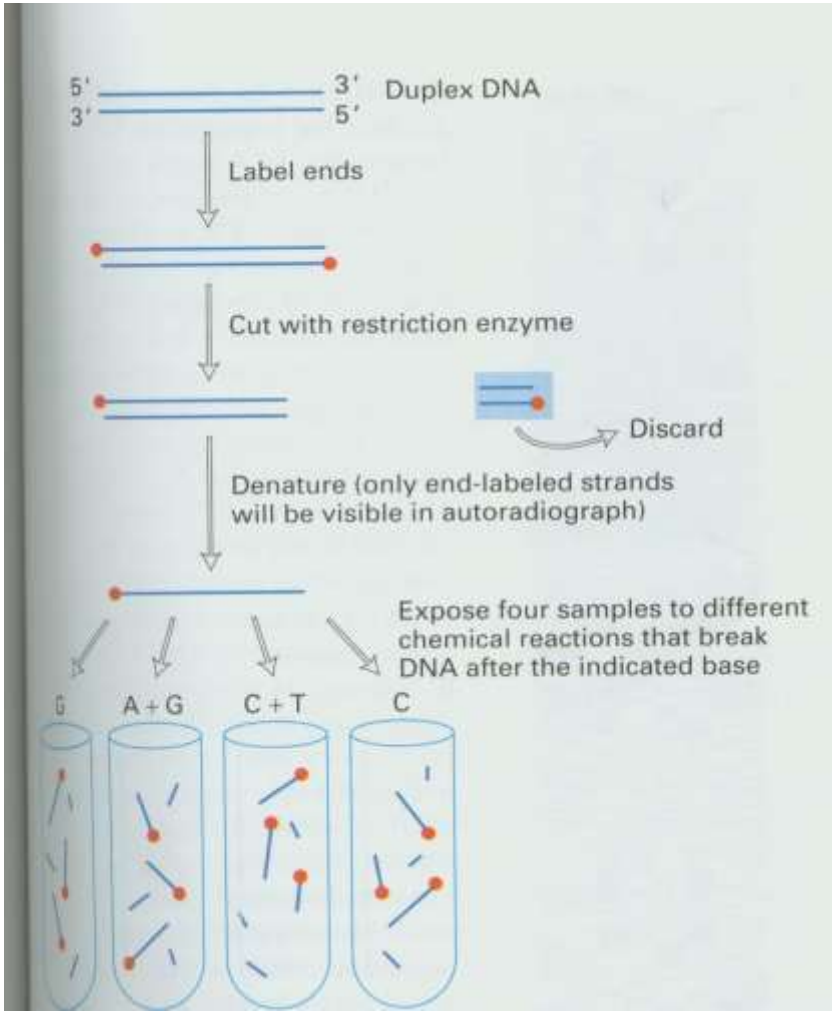
3. The fragment of DNA with one end labeled is denatured.

4. Four identical samples of these end-labeled DNA restriction fragments are subjected to chemical

cleavage at different chemical nucleotides.

5. There are four specific sets of chemical reactions that selectively cut the DNA backbone at G, A+G, C+T, or C residues.

- G only: Dimethyl sulphate(DMS)
  and piperidine
- A+G : DMS,and formamide piperidine
- C+T : Hydrazine, piperidine
- C only : Hydrazine, alkali or NaCl
  piperidine

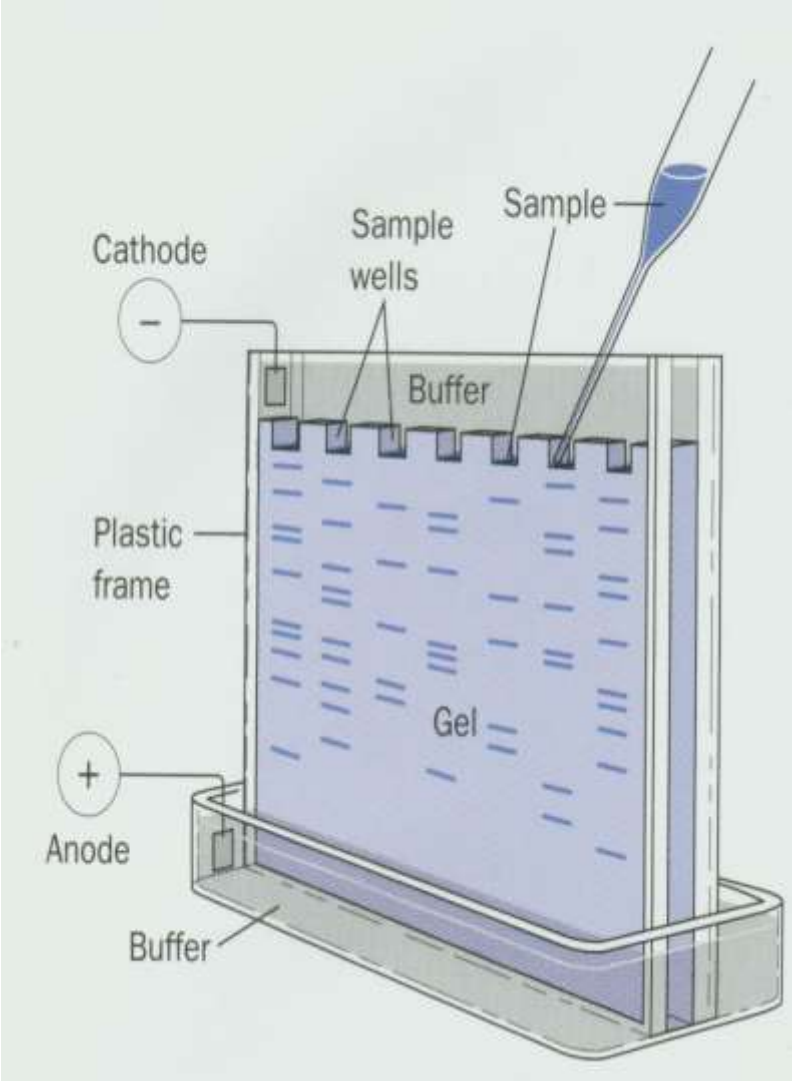


6. For each labeled chain to be broken only once, the reactions are controlled.

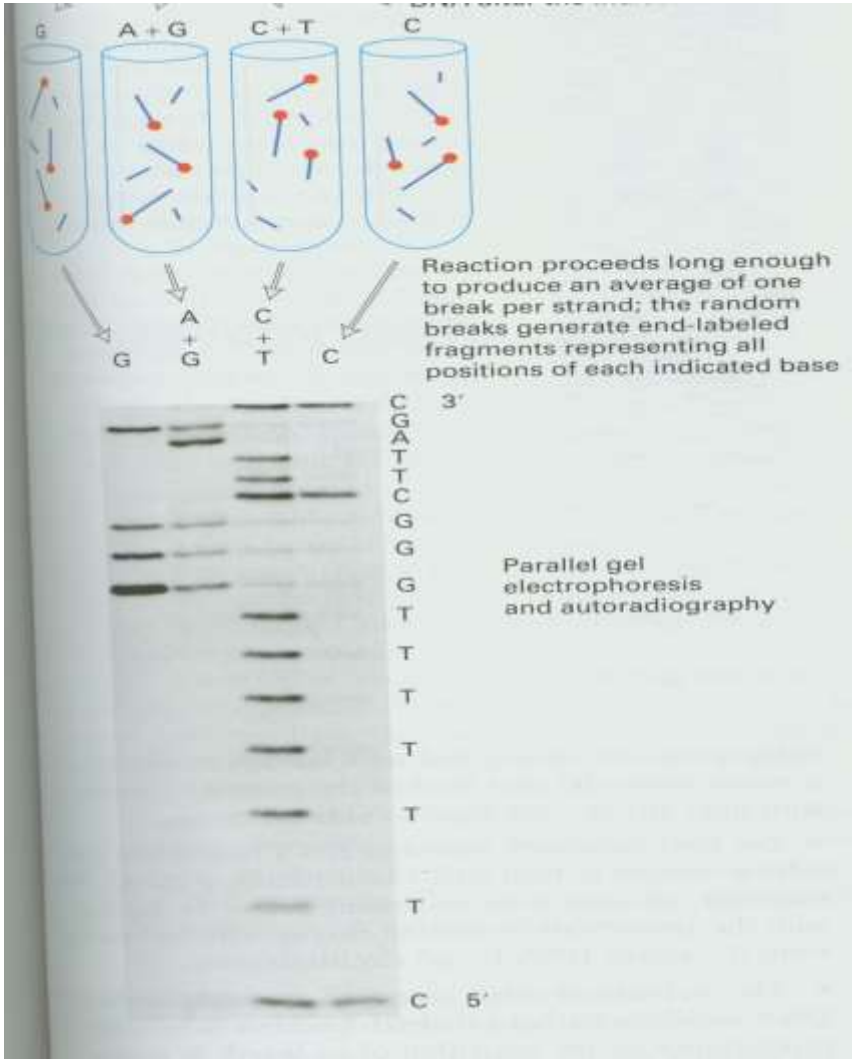
7. The labeled subfragments created by the four reactions have

- the ^{32}P label at one end and
- the chemical cleavage point at the other end.

8. The reaction products are separated by polyacrylamide gel electrophoresis which is based on size. Smallest fragment goes fastest.



9. The labeled fragments in the gel are visualized by autoradiography (x-ray).
10. The sequence is read from bottom to top of the gel.



EXAMPLE OF DNA SEQUENCING BY CHEMICAL METHOD

CHEMICAL CLEAVAGE OF A DNA SAMPLE AT C BASES

End-labelled DNA sample

^{32}P-A-p-T-p-T-p-G-p-C-p-G-p-C-p-T-p-G-p-C-p-A-p-C-p-G-p-C-p-T

^{32}P _____

^{32}P _____

^{32}P _____

^{32}P _____

^{32}P _____

End-labelled DNA fragments

AUTORADIOGRAM OF SAMPLE
MAXAM-GILBERT SEQUENCING GEL

G	A+G	C+T	C	SEQUENCE (END)
		—	—	C (3')
—	—			G
	—			A
		—		T
		—		T
		—		T
		—	—	C
—	—			G
—	—			G
	—			A
		—		T
		—	—	C
	—			A
	—			A (5')

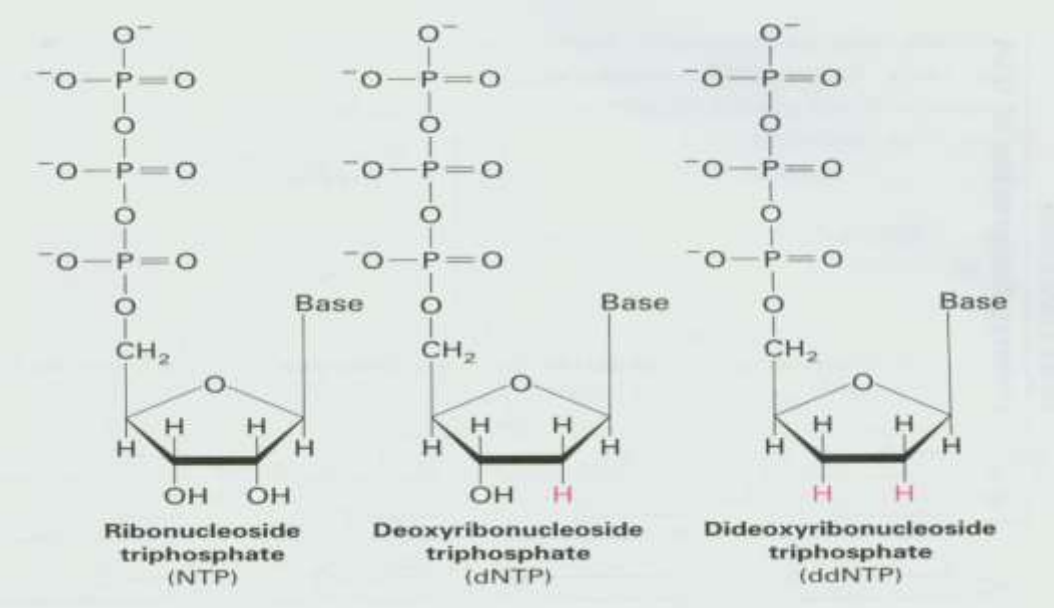
ADVANTAGES AND DISADVANTAGES

- No premature termination due to DNA sequencing. So, no problem with polymerase to synthesize DNA.
- Stretches of DNA can be sequenced which can not be done with enzymatic method.

- Not widely used.
- Use of radioactivity and toxic chemicals.

CHAIN TERMINATION METHOD

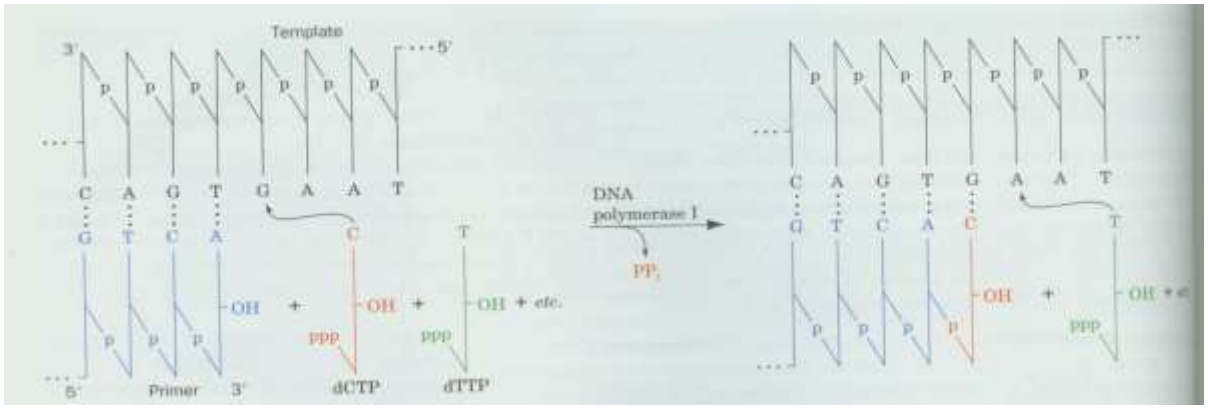
- This method uses single-stranded DNA.
- Also known as **dideoxy sequencing** method because it involves the use of analogue of normal nucleotide 2',3'-dideoxynucleoside triphosphates (ddNTPs). These are chain terminating nucleotides lacking 3'-OH ends.
- This method is based upon the incorporation of ddNTPs into a growing DNA strand to stop chain elongation.



Structure of NTP, dNTP, and ddNTP

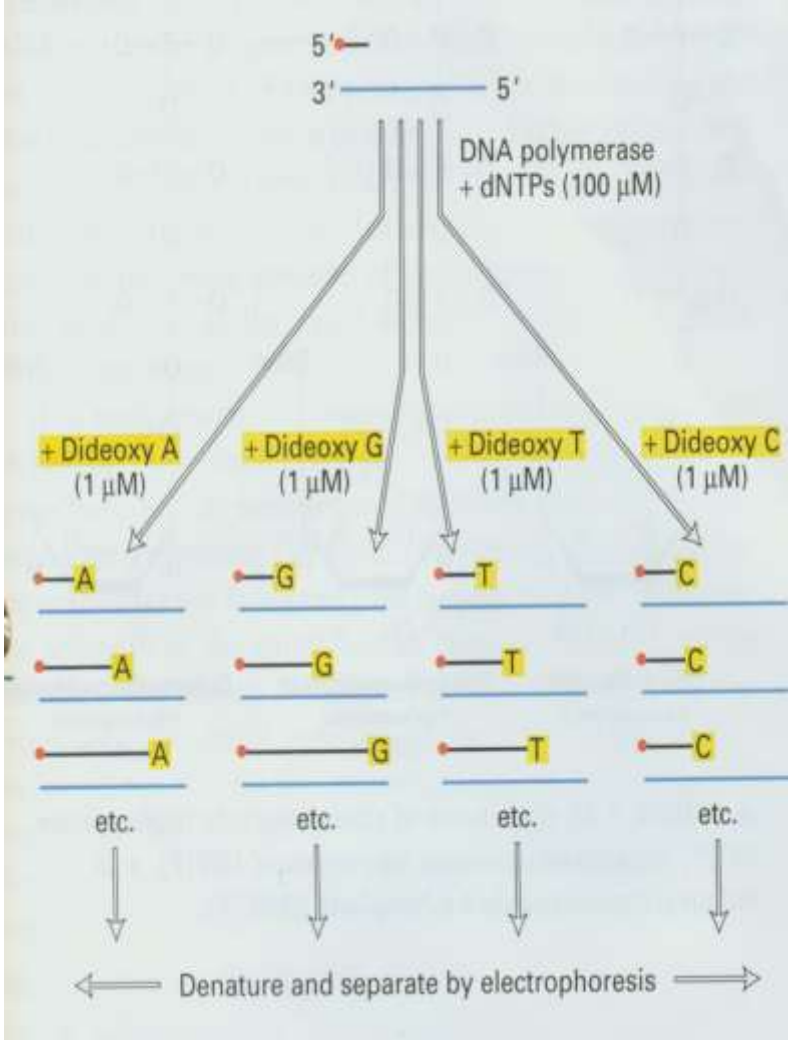
STAGES:

1. The DNA to be sequenced is extracted from phage or E. coli for sequencing purpose.
2. A synthetic 5'-end-labeled oligodeoxynucleotide is used as the primer.
3. The template DNA is hybridized to the primer.
4. The primer elongation is performed in four separate polymerization reaction mixtures. Each mixture contains
   - 4 normal deoxynucleotides (dNTPs) in higher concentration and
   - a low concentration of the each of the 4 ddNTPs.
5. There is initiation of DNA synthesis by adding enzyme DNA polymerase since the enzyme cannot distinguish between the normal nucleotides and their analogues.



Action of DNA polymerase I

6. The strand synthesis continues until a ddNTP is added. The chain elongation ceases on the incorporation of a ddNTP because it lacks a 3'-OH group which prevents addition of the next nucleotide.
7. There is a result of mixture of terminated fragments, all of different lengths.
8. Denature DNA fragments.
9. Each of the four mixtures are run together on a polyacrylamide gel for electrophoresis.

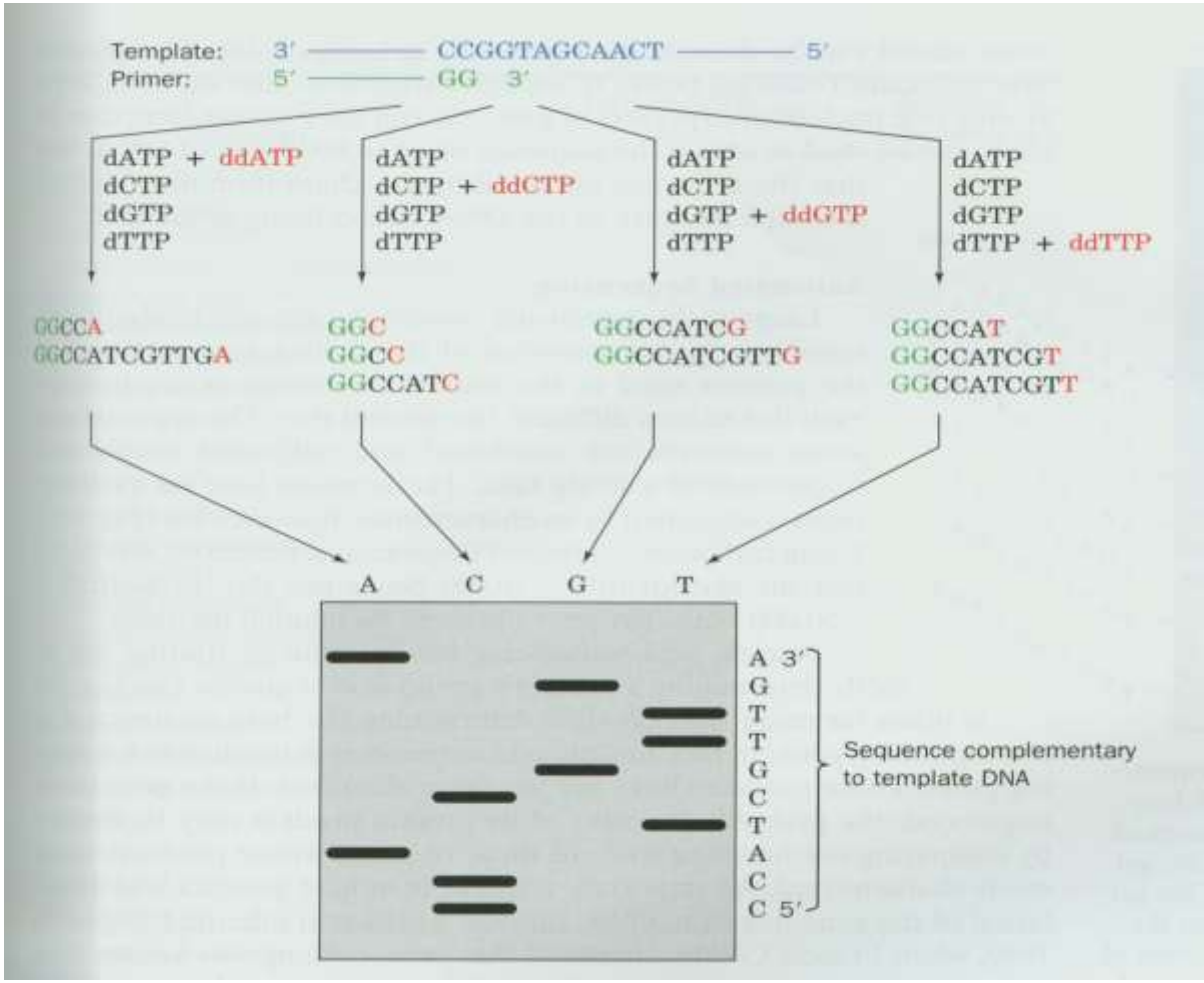


Sanger method

10. The separated fragments are

then visualized by autography.

11. From the position of the bands of the resulting autoradiogram, the sequence of the original DNA template strand can be read directly.



Chain termination method

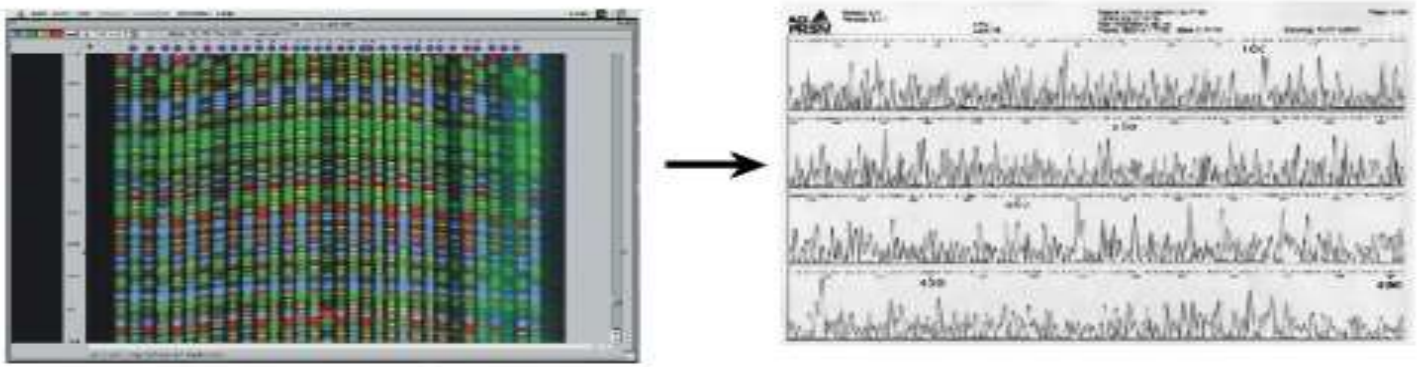
ADVANTAGES AND DISADVANTAGES

- Most popular method.
- Simpler and quicker allowing large output. Within an hour the primer-annealing and sequencing reactions can be completed.
- Yielding of poor results owing to secondary structure in the DNA as sometimes DNA polymerases terminate chain elongation prematurely.
- The sequence is obtained not from the original DNA molecule but from an enzymatic copy. So, there is a chance of incorporation of wrong bases.

IMPROVED APPROACHES AND AUTOMATED DNA SEQUENCING

- Updated version of Sanger method
- Fluorescence detection with lasers
- Cycle sequencing
- Shotgun sequencing

Automated procedure for DNA sequencing



A computer read-out of the gel generates a "false color" image where each color corresponds to a base. Then the intensities are translated into peaks that represent the sequence.

CYCLE SEQUENCING

- There are two basic differences between cycle sequencing and PCR amplification:
  - The presence of only one primer in the cycle-sequencing reaction used to prime synthesis of one strand of the DNA
  - The presence of dideoxynucleotide triphosphates in the sequencing reactions that create the base-specific terminations required.
- The result of the temperature cycling is linear amplification of the sequencing product leading to an increase in the signal generated during the sequencing reaction when compared with standard sequencing protocols.

CYCLE SEQUENCING

- Cycling the sequencing reactions results in several advantages

(1) The amount of template necessary for the sequencing reaction is greatly reduced

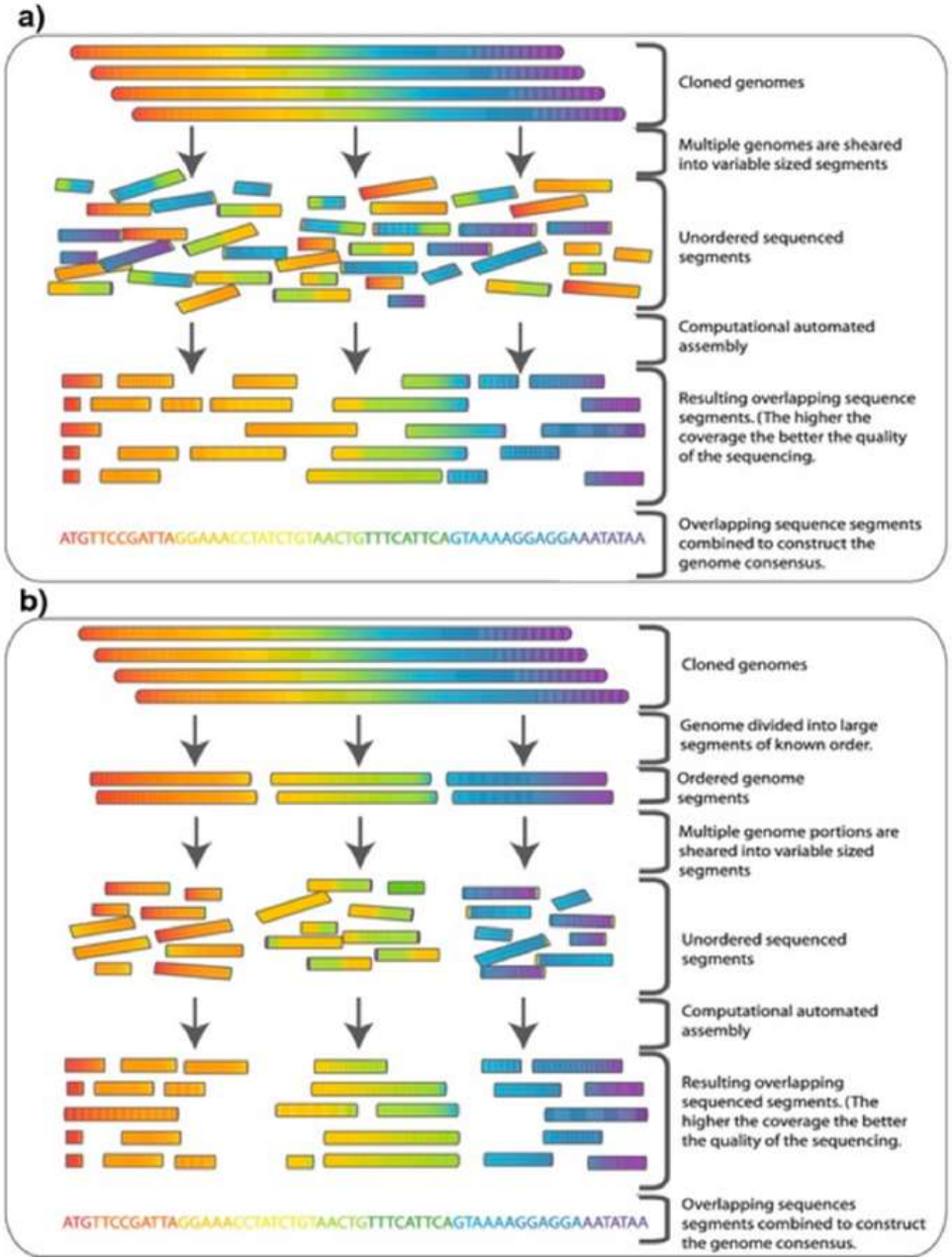
(2) because smaller amounts of template are added, fewer impurities are introduced, meaning less template preparation is required; and

(3) The high temperature at which the sequencing reactions are run and the multiple heat-denaturation steps allow double- stranded templates such as plasmids, cosmids, X DNA, and PCR products to be sequenced reliably without a separate denaturation step

SHOTGUN SEQUENCING

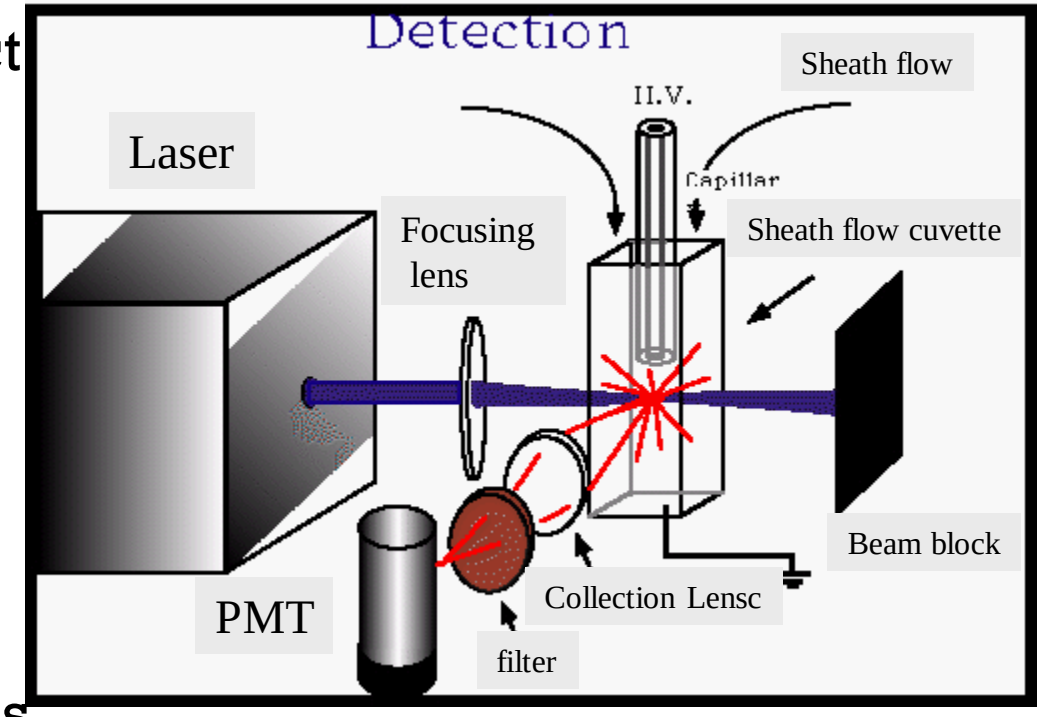
- is a method used for sequencing long DNA strands
- DNA is broken up randomly into numerous small segments, which are sequenced using the chain termination method to obtain *reads*.
- Multiple overlapping reads for the target DNA are obtained by performing several rounds of this fragmentation and sequencing.
- Computer programs then use the overlapping ends of different reads to assemble them into a continuous sequence

SHOTGUN SEQUENCING



High-throughput sequencing: Capillary electrophoresis

The human genome project has spurred an effort to develop faster, higher throughput, and less expensive technologies for DNA sequencing. Capillary electrophoresis (CE) separation has many advantages over slab gel separations. CE separations are faster and are capable of producing greater resolution. CE instruments can use tens and even hundreds of capillaries simultaneously. The figure show a simple CE setup where the fluorescently-labeled DNA is detected as it exits the capillary.

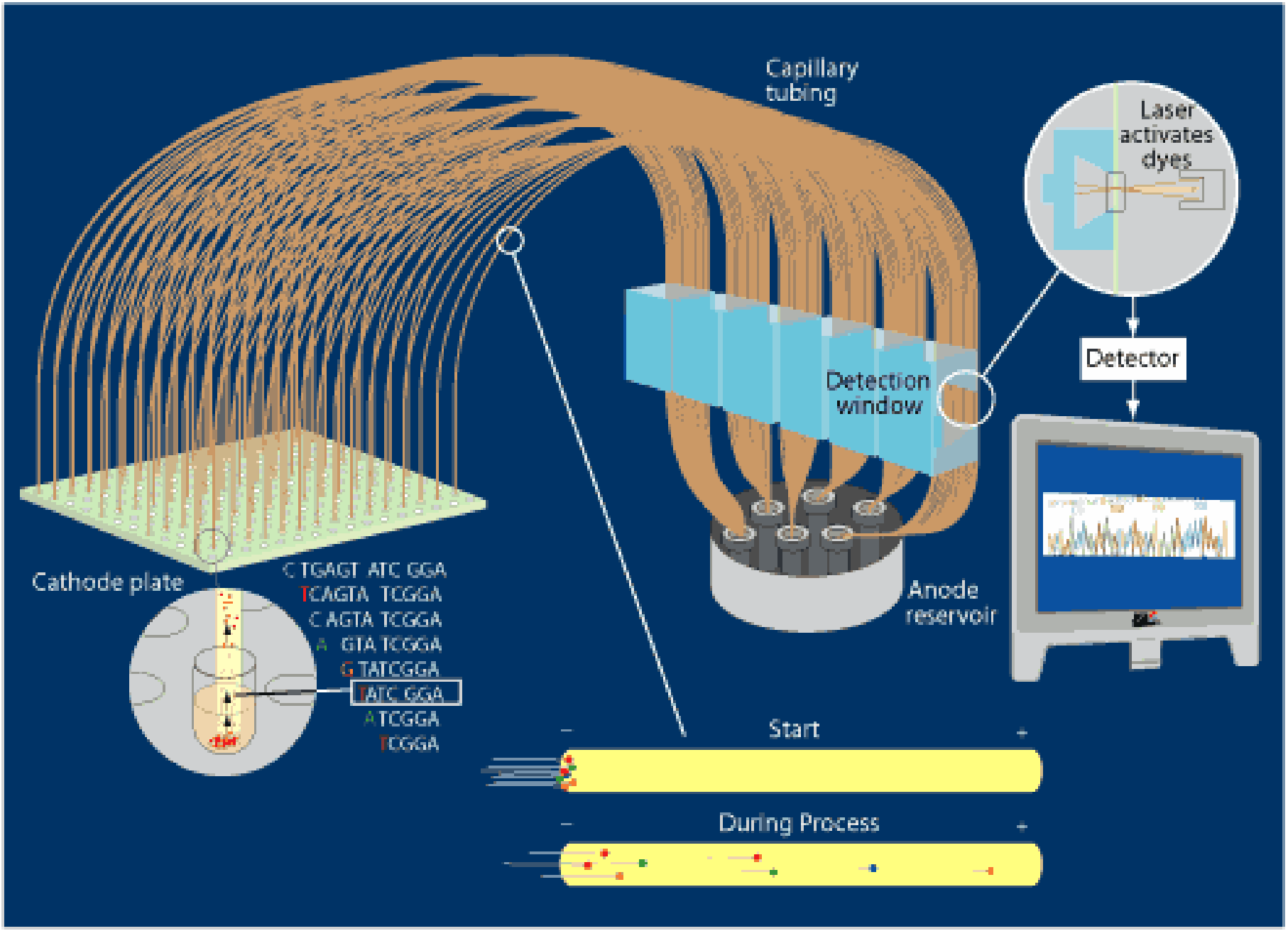


Sieving matrix for CE

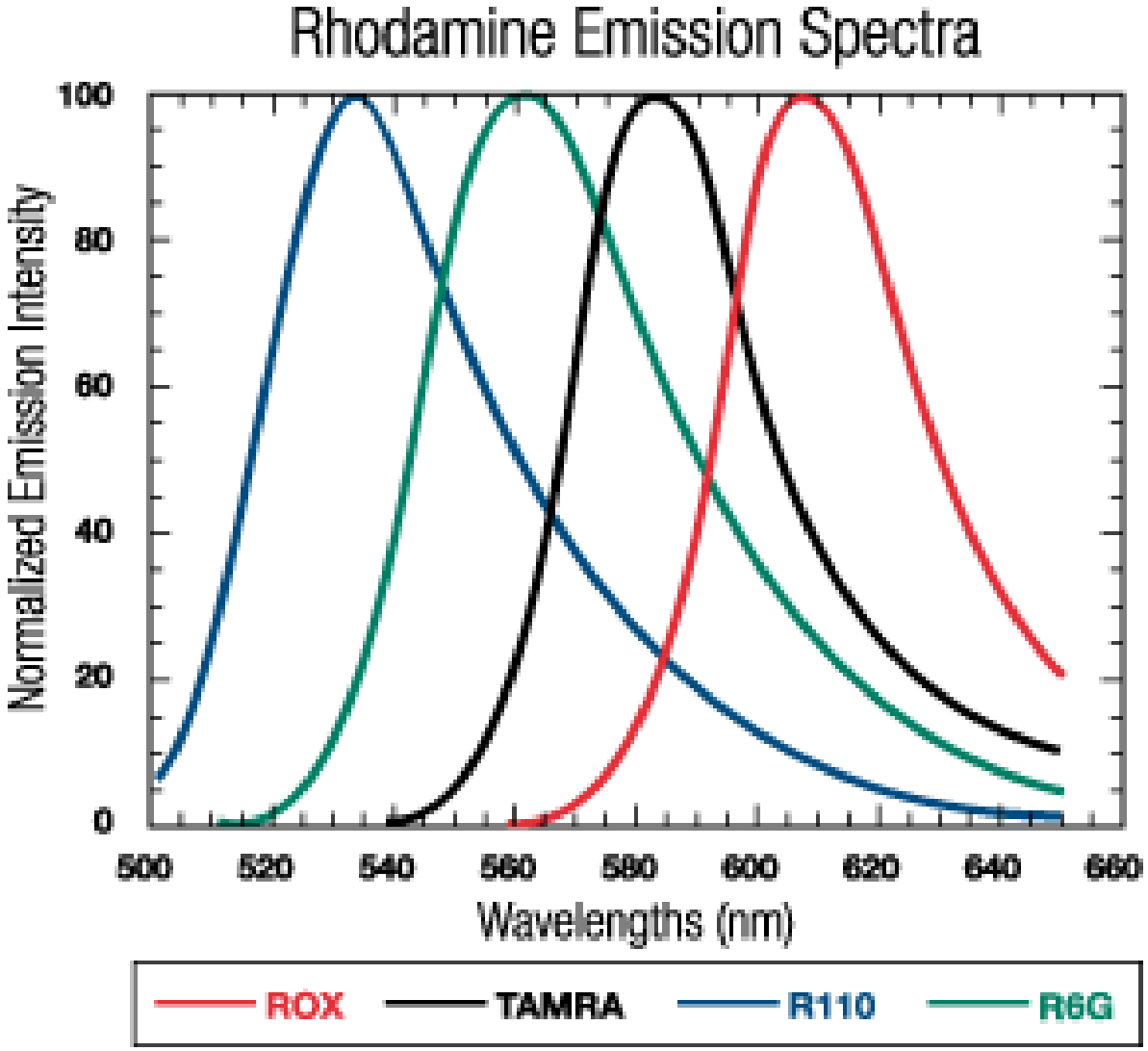
•To separate DNA fragments of different sizes the capillary needs to be filled with sieving matrix, such as **linear polyacrylamide** (acrylamide polymerized without bis-acrylamide).

•This material is not rigid like a cross-linked gel but looks much like glycerol. With a little bit of effort it can be pumped in and out of the capillaries.

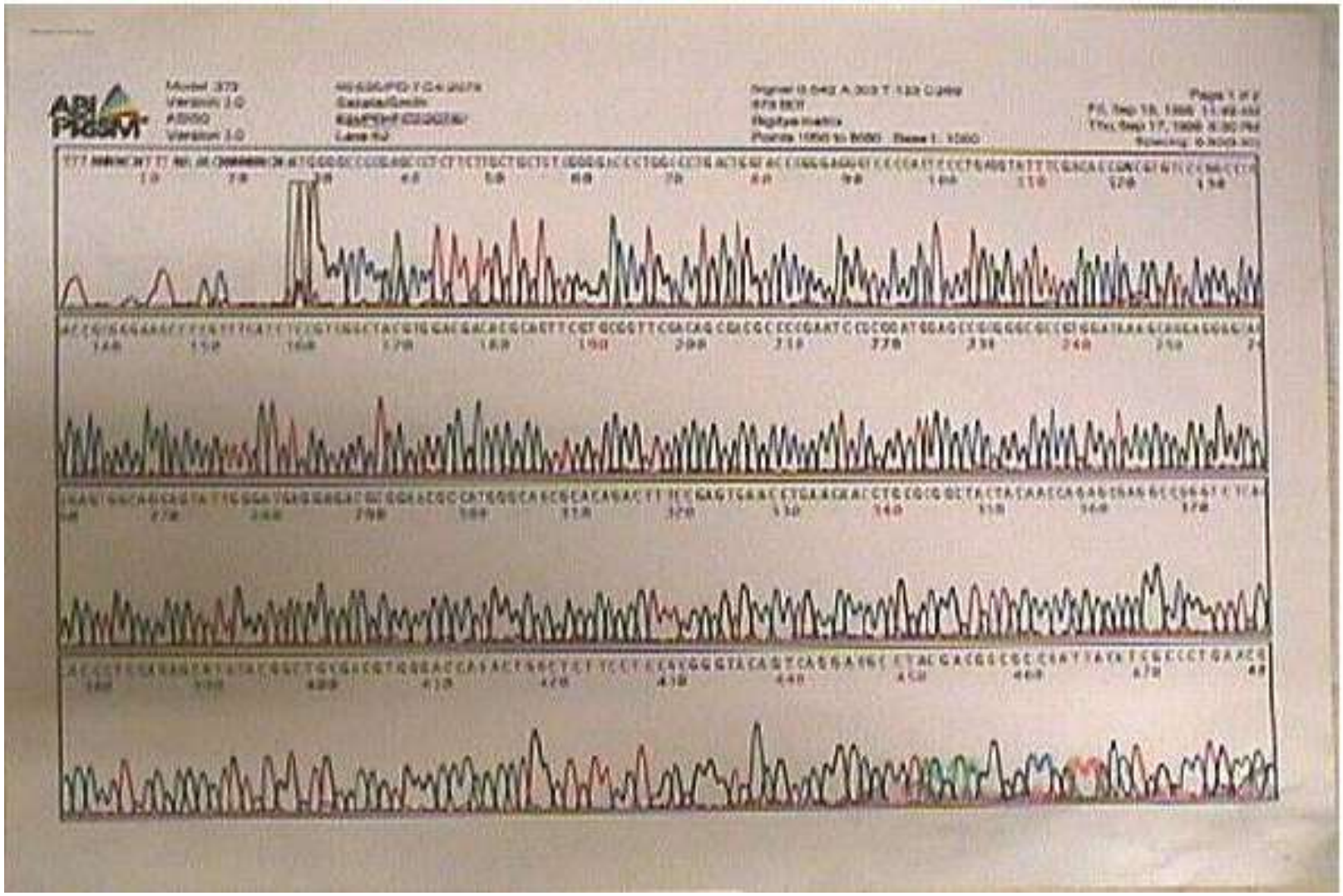
•To simulate the separation characteristics of an agarose gel one can use **hydroxyethylcellulose**. It is not much more viscous than water and can easily be pumped into the capillaries.


Capillary tubing
Laser activates dyes
Detector
Detection window
Anode reservoir
Cathode plate
C TGAGT ATC GGA
TCAGTA TCGGA
C AGTA TCGGA
A GTA TCGGA
G TATCGGA
TATC GGA
A TCGGA
TCGGA
Start
During Process

Fluorescent end labeling of DNA



DNA SEQUENCING OUTPUT



What is the function of the sequenced gene?

Classical methods:

- mutate gene, characterize phenotype for clues to function (genetics)

- purify protein product, characterize *in vitro* (biochemistry)

Comparison to previously characterized genes:

- genes sequences that have high sequence similarity usually have similar functions

- if your gene has been previously characterized (using classical methods) by someone else, you want to know right away! (avoid duplication of labor)

NCBI

NCBI home page -Go to www.ncbi.nlm.nih.gov for the following pages

Pubmed: search tool for literature--search by author, subject, title words, etc.

All databases: "a retrieval system for searching several linked databases"

BLAST: Basic Local Alignment Sequence Tool

OMIM: Online Mendelian Inheritance in Man

Books: many online textbooks available

Tax Browser: A taxonomic organization of organisms and their genomes

Structure: Clearinghouse for solved molecular structures

What does BLAST do?

1) Searches chosen sequence database and identifies sequences with similarity to test sequence
2) Ranks similar sequences by degree of homology (E value)
3) Illustrates alignment between test sequence and similar sequences

Alignment of sequences:

The principle: two homologous sequences derived from the same ancestral sequence will have at least some identical (similar) amino acid residues

Fraction of identical amino acids is called "percent identity"

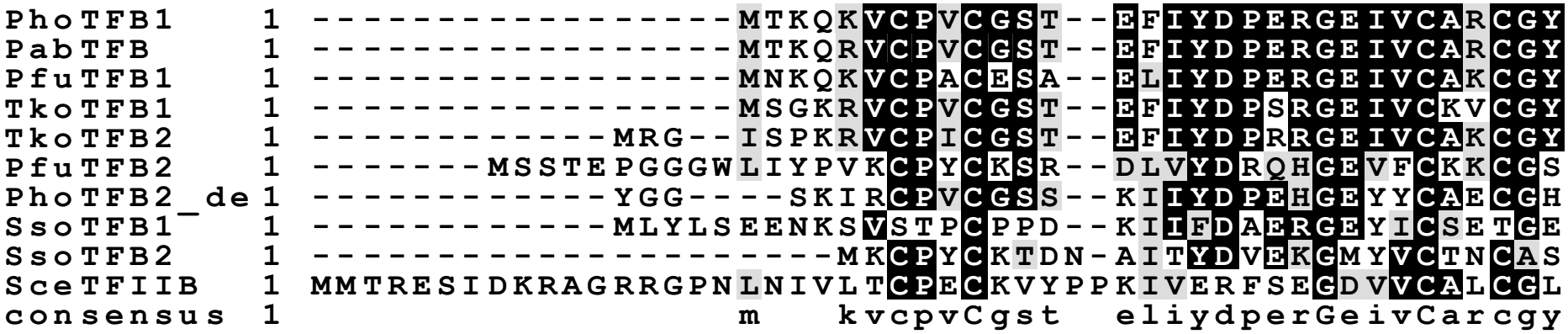
Similar amino acids: some amino acids have similar physical/chemical properties, and more likely to substitute for each other-these give specific similarity scores in alignments

Gaps in similar/homologous sequences are rare, and are given penalty scores

Homology of proteins

Homology: similarity of biological structure, physiology, and development based on genetic inheritance

Homologous proteins: statistically similar sequence, therefore similar functions (often, but not always)



Alignment of TFB and TFIIB sequences
TFB stands for archaeal transcription factor

Translating the DNA sequence

The order of amino acids in any protein is specified by the order of nucleotide bases in the DNA.
Each amino acid is coded by the particular sequence of three bases.
To convert a DNA sequence

First, find the starting codon. The starting codon is always the codon for the amino acid methionine. This codon is AUG in the RNA (or ATG in the DNA):

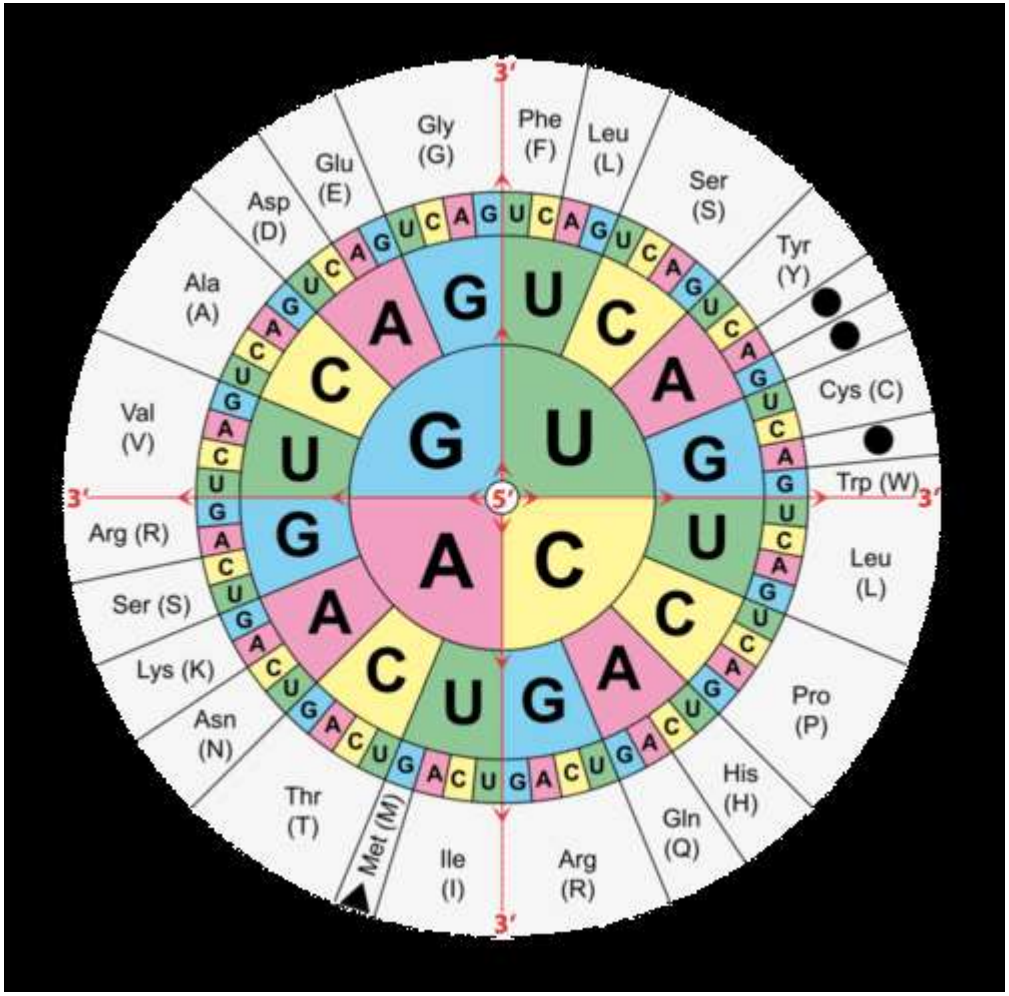
GCGCGGGUCCGGGCAUGAAGCUGGGCCGGGCCGUGC....
Met

In this particular example the next codon is AAG. The first base (5'end) is A, so that selects the 3rd major row of the table. The second base (middle base) is A, so that selects the 3rd column of the table. The last base of the codon is G, selecting the last line in the block of four.

The codon table

5'-Base	Middle Base				3'-Base
	U(=T)	C	A	G	
U(=T)	Phe	Ser	Tyr	Cys	U(=T)
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Term	Term	A
	Leu	Ser	Term	Trp	G
C	Leu	Pro	His	Arg	U(=T)
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U(=T)
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U(=T)
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The codon table



Translating the DNA sequence

This entry AAG in the table is Lysine (Lys).
Therefore the second amino acid is Lysine.

The first few residues, and their DNA sequence, are as follows (color coded to indicate the correct location in the codon table):

Met Lys Leu Gly Arg ... ...
AUG AAG CUG GGC CGG GCC GUG C..

This procedure is exactly what cells do when they synthesize proteins based on the mRNA sequence. The process of translation in cells occurs in a large complex called the ribosome.

HUMAN GENOME PROJECT (HGP)

HGP is a national effort to sequence and analyze the human genome which is a very complex system consisting of 50,000 to 100,000 genes. These genes are located on 23 base pairs of chromosome. The complete sequence was complete in 2005.

Some reasons for studying Human genome:

- Better medical practice
- High-quality diagnosis of diseases
- Understanding of evolution fully
- Improvement in biological research and forensic science
- Improvement in agriculture etc.

The latest research on HGP are

- Pulsed electrophoresis
- Fluorescence microscopy
- 2D gel electrophoresis
- gtc double-stranded subclone inserts