Fundamentos de Biologia Molecular

Curso de Licenciatura em Biologia
2º Ano, 1º Semestre
Ano Letivo 2021/2022

Componente Teórico-Prática

Ciências ULisboa
Faculdade de Ciências da Universidade de Lisboa

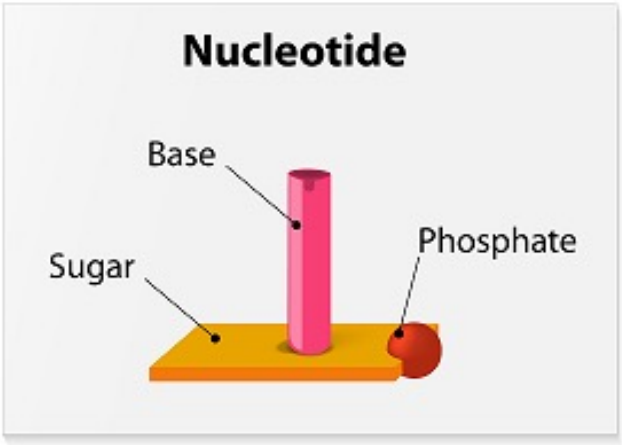
Docente Responsável: ***Rita Zilhão***
Docente TPs: ***Andreia Figueiredo***

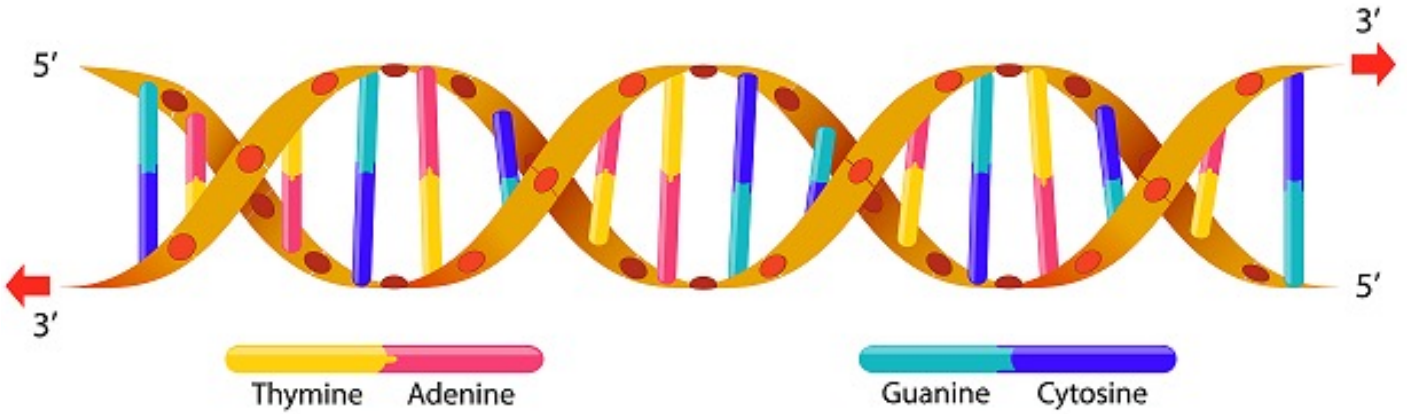
TP3: DNA Sequencing

- Basic principle of Sanger sequencing: DNA structure
- Nobel prizes
- Dideoxy-terminating DNA/Sanger sequencing
- Overview of Sanger sequencing steps
- Technical advances
- Sequencing data analysis

TP3: DNA Sequencing

DNA structure





TP2: Polymerase Chain Reaction (PCR)

Frederick Sanger

- After his Ph.D. in 1943, Sanger started working for A. C. Chibnall, on identifying the free amino groups in insulin. In the course of identifying the amino groups, Sanger figured out ways to order the amino acids. He was the first person to obtain a protein sequence. By doing so, Sanger proved that proteins were ordered molecules and by analogy, the genes and DNA that make these proteins should have an order or sequence as well – first Nobel Prize in 1958

- Solving the problem of DNA sequencing became a natural extension of his work in protein sequencing. Sanger initially investigated ways to sequence RNA because it was smaller. Eventually, this led to techniques that were applicable to DNA and finally to the **dideoxy method most commonly used in sequencing reactions today**. Sanger won a second Nobel Prize for Chemistry in 1980 sharing it with Walter Gilbert, for their contributions concerning the determination of base sequences in nucleic acids, and Paul Berg for his work on recombinant DNA.

TP2: Polymerase Chain Reaction (PCR)

Paul Berg

▪An organism's genome is stored in the form of long rows of building blocks, known as nucleotides, which form DNA molecules. ***An organism's genome can be mapped by establishing the order of the nucleotides within the DNA molecule***. In 1976, Allan Maxam and Walter ***Gilbert developed a method by which the ends of the DNA molecule could be marked using radioactive substances***. After undergoing treatment with small amounts of chemicals that react with specific nucleotides, DNA fragments of varying lengths can be obtained. After undergoing what is known as electrophoresis, the nucleotide sequences in a DNA sample can be identified.

Walter Gilbert

▪DNA carries organisms' genomes and also determines their vital processes. The ability to artificially manipulate DNA opens the way to creating organisms with new characteristics. In conjunction with his studies of the tumor virus SV40, in 1972, Paul Berg succeeded in inserting DNA from a bacterium into the virus' DNA. He thereby **created the first DNA molecule made of parts from different organisms-"hybrid DNA" or "recombinant DNA"**.

TP3: DNA Sequencing

Frederick Sanger

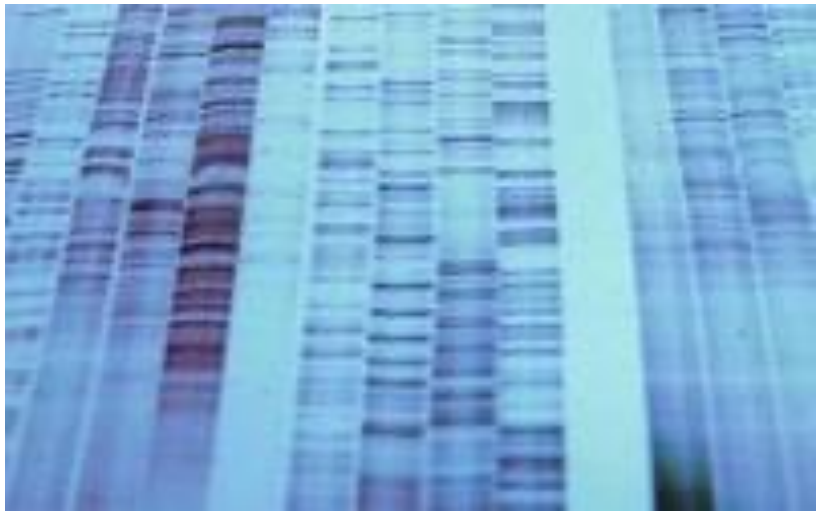
Paul Berg

Walter Gilbert

The Nobel Prize in Chemistry 1980 was divided, one half awarded to Paul Berg ***"for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant-DNA"***, the other half jointly to Walter Gilbert and Frederick Sanger ***"for their contributions concerning the determination of base sequences in nucleic acids"*.**

TP3: DNA Sequencing

Dideoxy-terminating DNA/Sanger sequencing concept

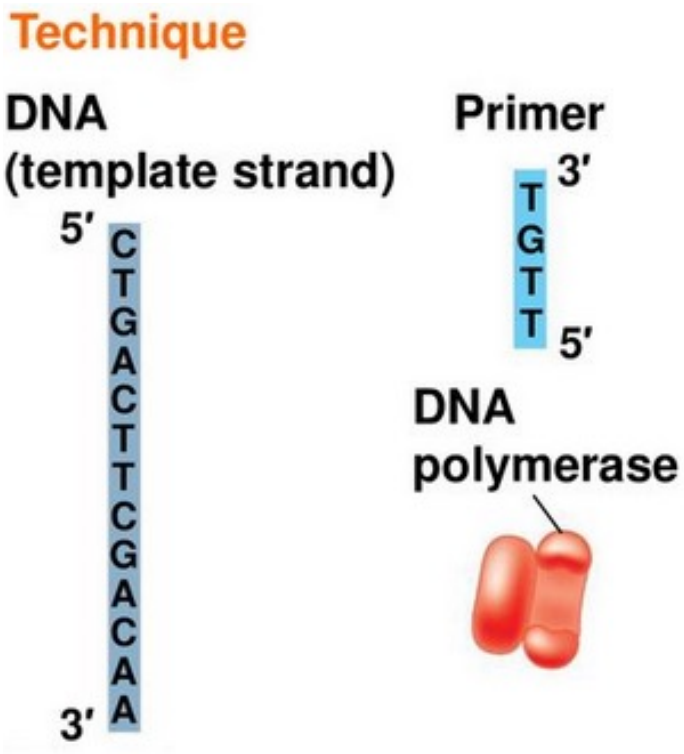


This method begins with the use of **special enzymes to synthesize fragments of DNA that terminate when a selected base appears in the stretch of DNA being sequenced**. These fragments are then sorted according to size by electrophoresis. Because of DNA's negative charge, the fragments move across the gel toward the positive electrode. The shorter the fragment, the faster it moves. Typically, each of the terminating bases within the collection of fragments is tagged with a **radioactive probe** for identification.

https://dnalc.cshl.edu/view/15479-Sanger-method-of-DNA-sequencing-3D-animation-with-narration.html

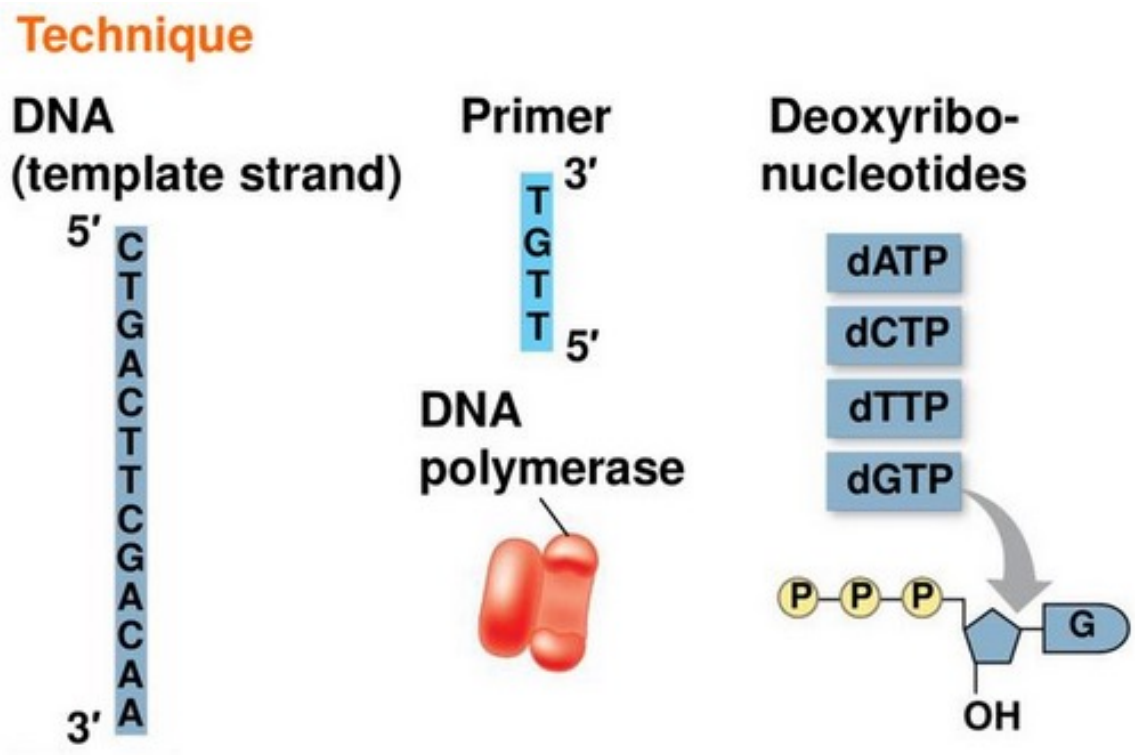
TP3: DNA Sequencing

Overview of Sanger sequencing steps



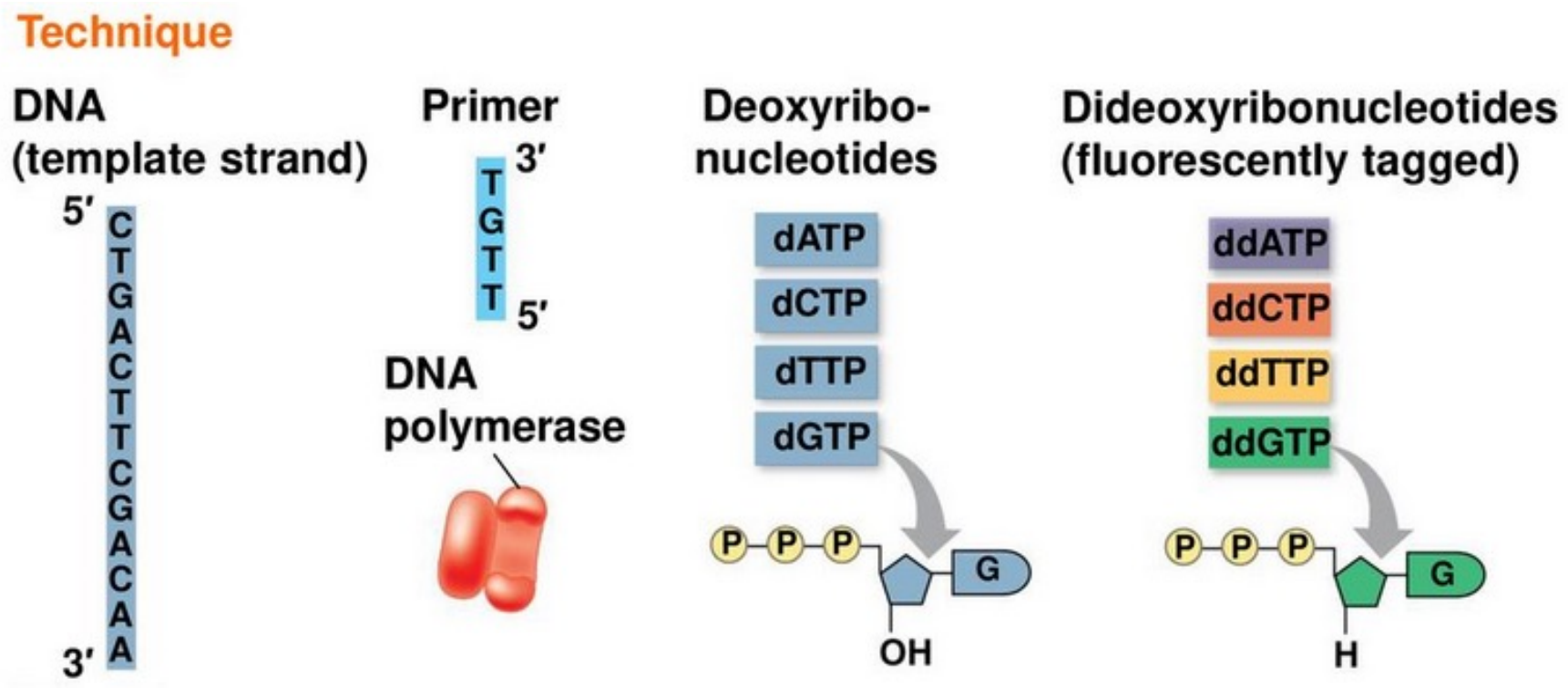
TP3: DNA Sequencing

Overview of Sanger sequencing steps



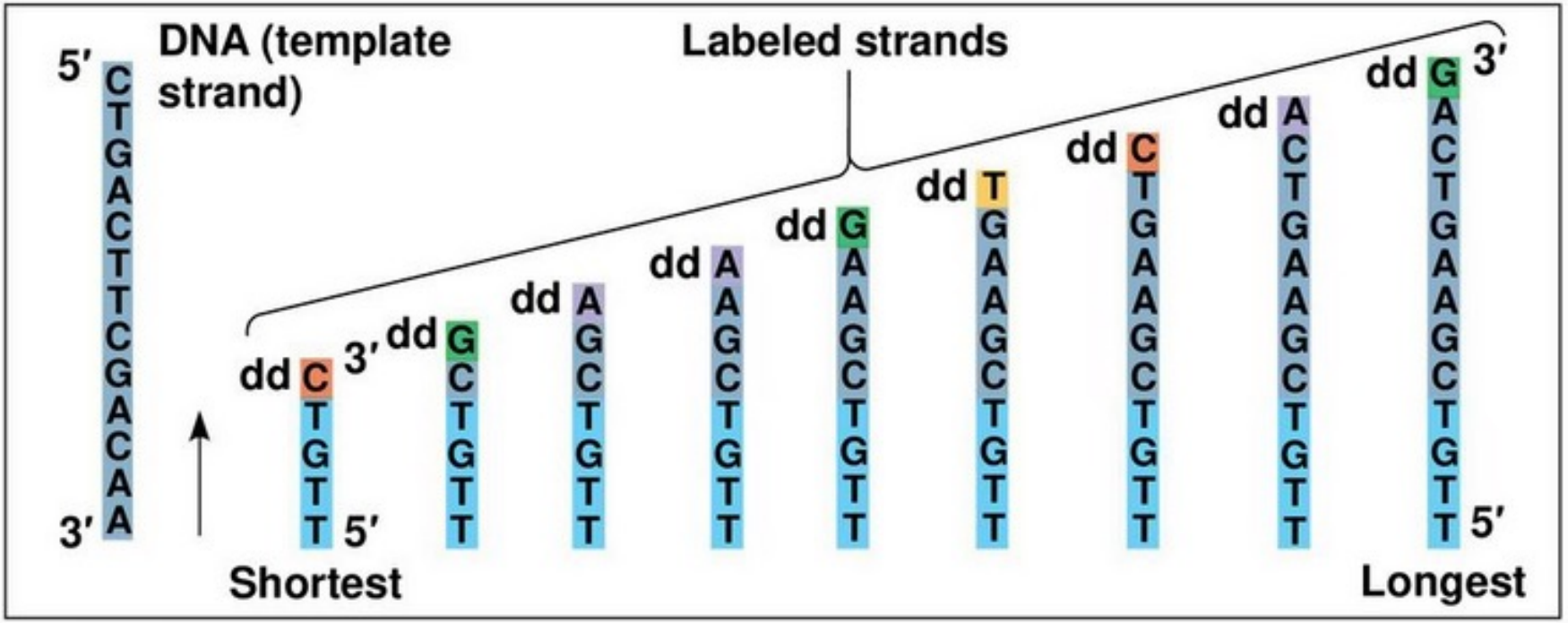
TP3: DNA Sequencing

Overview of Sanger sequencing steps



TP3: DNA Sequencing

Technique



TP3: DNA Sequencing

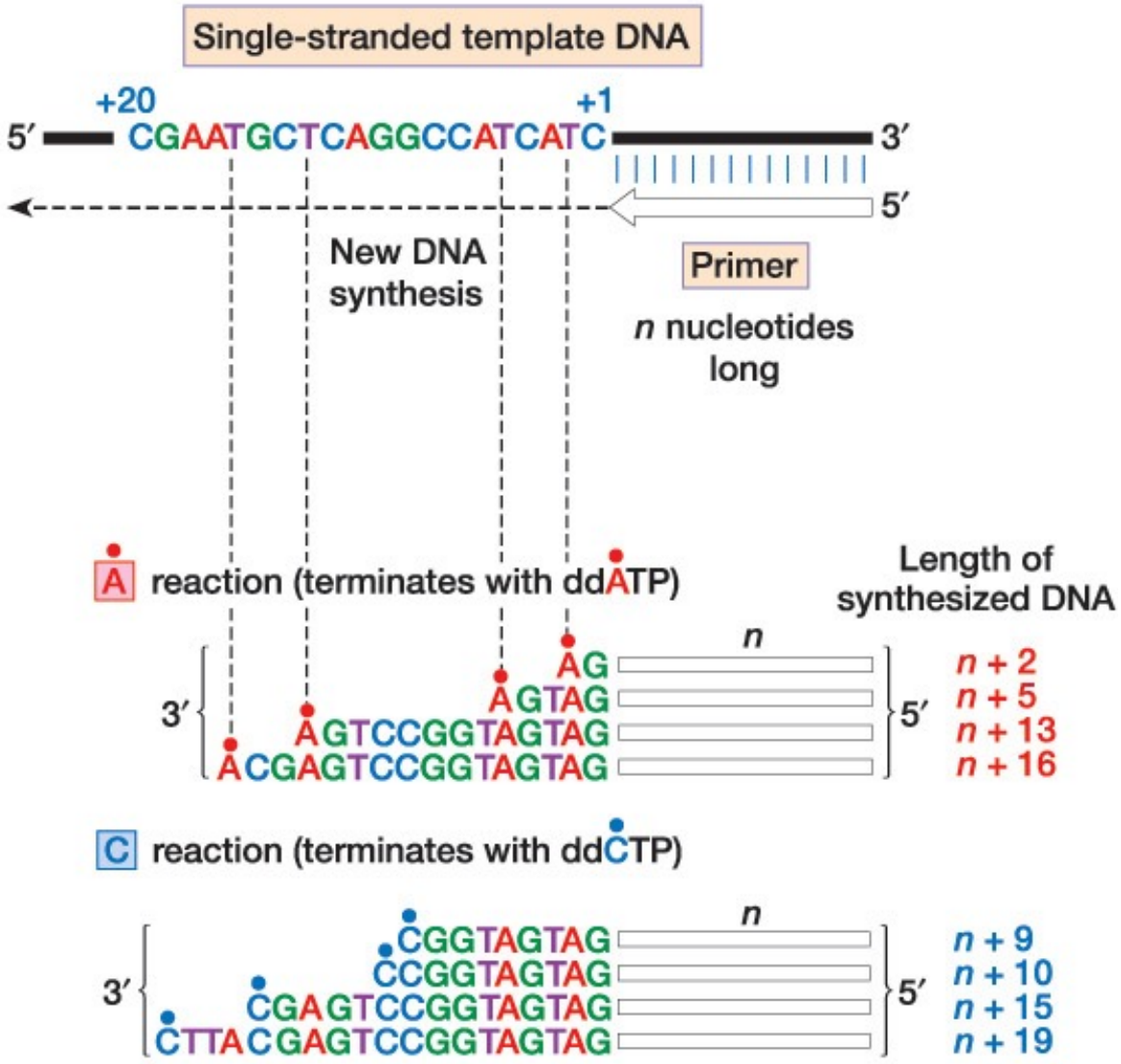
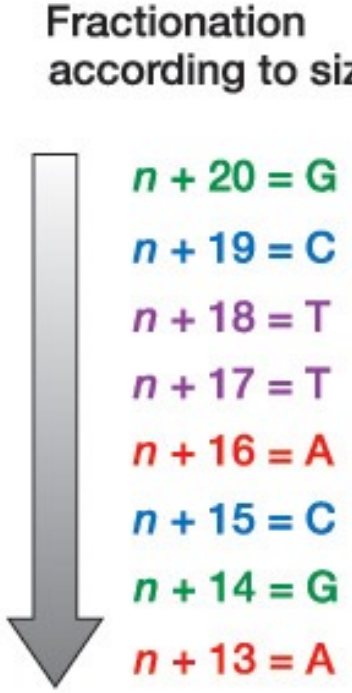


Figure 7-2 part 1 of 2 Human Molecular Genetics, 3/e. (© Garland Science 2004)



.n + 9 = C
.n + 5 = A
.n + 2 = A
.

TP3: DNA Sequencing

Overview of Sanger sequencing steps

-first denaturate DNA – separation of double chain
-Anneal the primer (1 primer that anneals to the region of interest)
-The DNA is placed into 4 different tubes, one for each nitrogenous base
-DNA polymerase and 4 deoxynucleotides are added to each tube (dNTPs)
-One type of dideoxynucleotides is added to each tube
-DNA polymerase extends the DNA sequence (from de primer 5'-3')
-No nucleotide can be added to the DNA chain once a dideoxynucleotides has been incorporated, so each fragment will end with a labeled nucleotide.
-The content of each tube is denaturated and separated by size by electrophoresis (polyacrylamide gel)
-So many sequences are synthesized that ddNTPs incorporation occurs in every sequence of the newly synthesized DNA sequence
-The further a specific strand has moved, the shorter it is – thus the position of the nucleotide that terminates that sequence can be determined by the distance travelled
-The order of nucleotides produced is a sequence (5'-3') that complements the original strand of DNA

TP3: DNA Sequencing

Sanger sequencing vs PCR

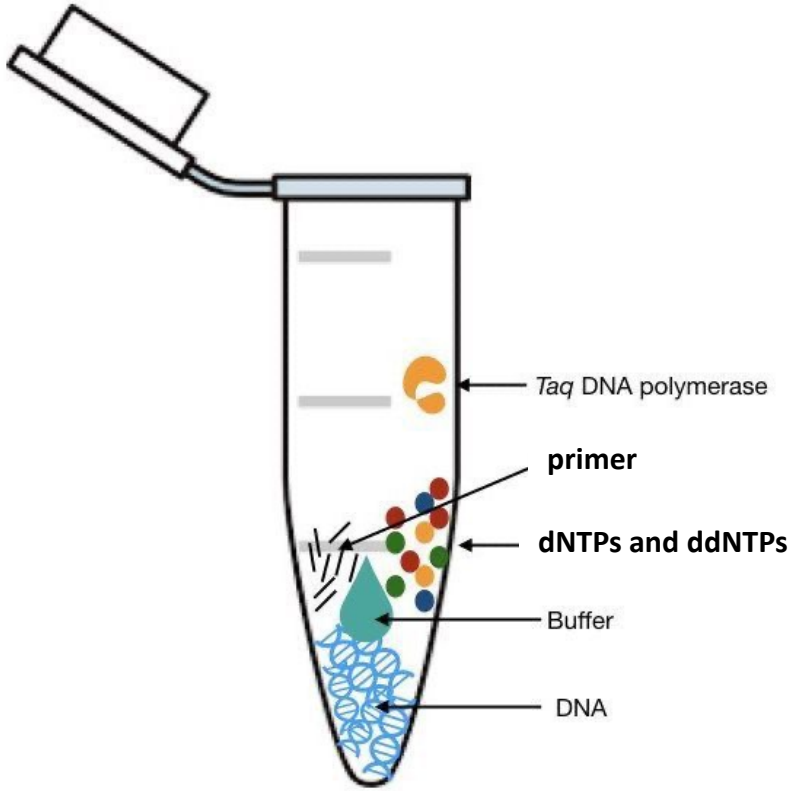
•**PCR is used to amplify DNA in its entirety**. While fragments of varying lengths may be produced by accident (e.g., the DNA polymerase might fall off), the goal is to duplicate the entire DNA sequence. To that end, the "ingredients" are the target DNA, nucleotides, DNA primer, and DNA polymerase (specifically Taq polymerase, which can survive the high temperatures required in PCR).

•**The goal of Sanger sequencing is to generate every possible length of DNA up to the full length of the target DNA**. That is why, in addition to the PCR starting materials, the **dideoxynucleotides** are necessary. Sanger sequencing and PCR can be brought together when generating the starting material for a Sanger sequencing protocol. PCR can be used to create many copies of the DNA that is to be sequenced. Having more than one template to work from makes the Sanger protocol more efficient.

TP3: DNA Sequencing

Dideoxy-terminating DNA sequencing reaction components

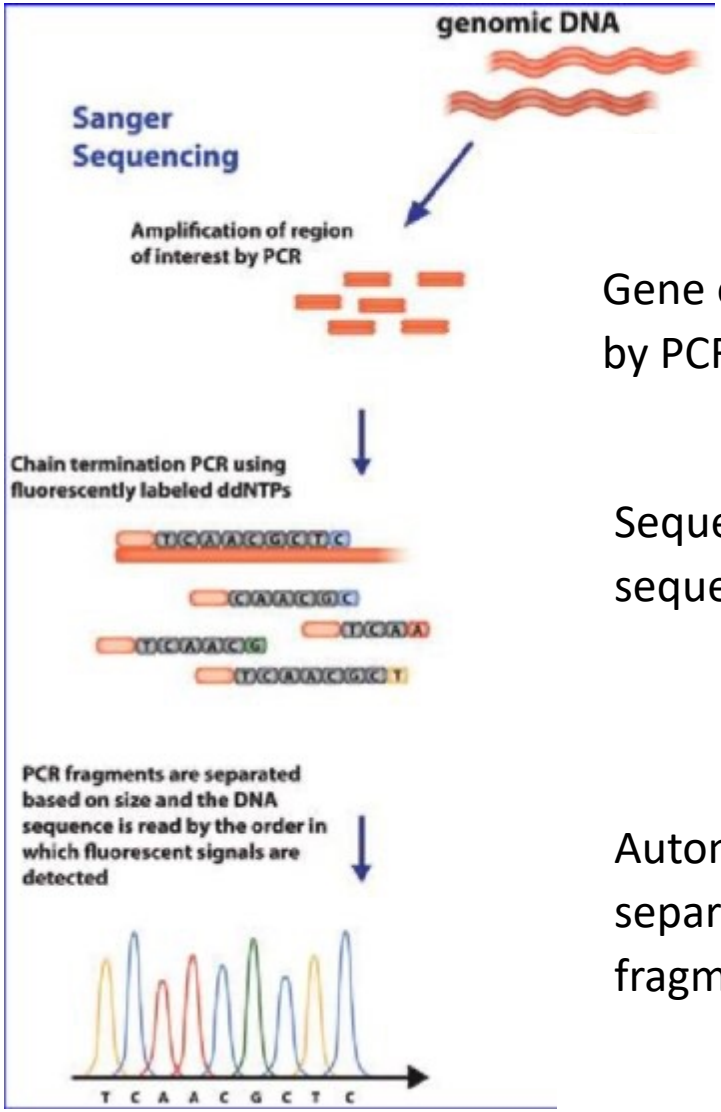
- **DNA template** to be sequenced
- One **specific primer** that binds to the template DNA and acts as a "starter" for the polymerase
- **nucleotides** (dATP, dTTP, dCTP, dGTP)
- **DNA polymerase** (proofreading activity)
- Dideoxy, or **chain-terminating**, versions of all four nucleotides (ddATP, ddTTP, ddCTP, ddGTP), each labeled (either radioactive label or fluorescent label with a different color of dye)



TP3: DNA Sequencing

DNA template to be sequenced

Specific DNA



Genomic DNA – billions of genes

Gene of interest – specifically amplified by PCR

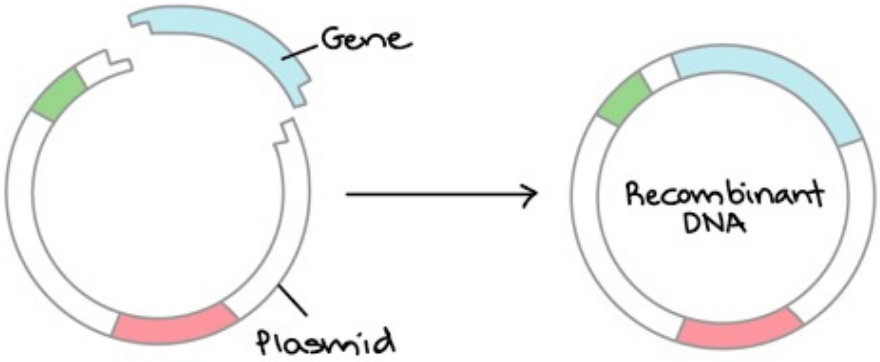
Sequencing reaction – Sanger sequencing (ddNTPs)

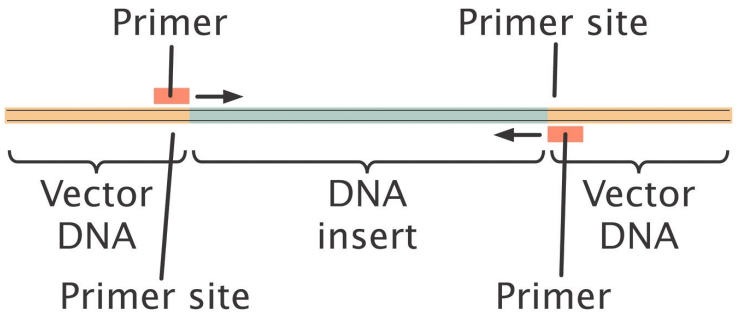
Automated sequencing by capillary separation of fluorescent labeled fragments

TP3: DNA Sequencing

DNA template to be sequenced

DNA cloned in a plasmid





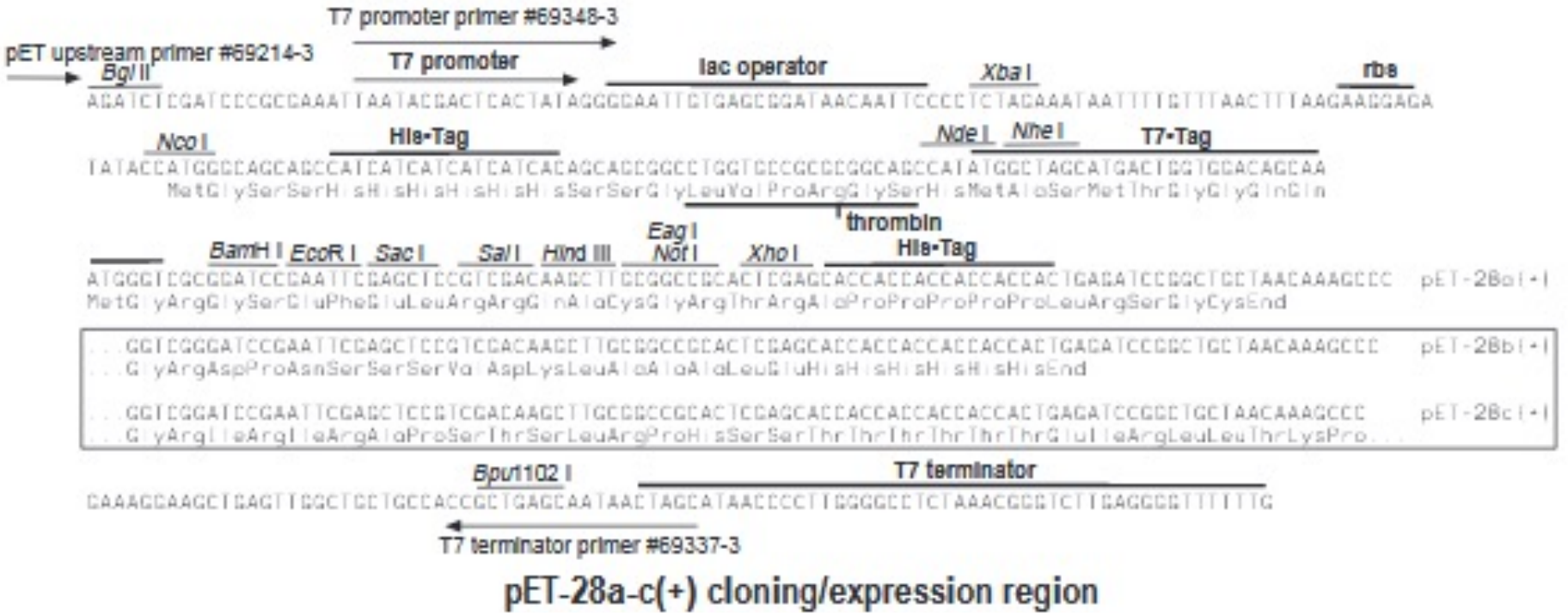
A **universal sequencing primer** can be used to sequence many different template DNAs (eg M13, T7 primers)

Vectors contain it on either side of the site where DNA will be inserted

TP3: DNA Sequencing

DNA template to be sequenced

DNA cloned in a plasmid



pET-28a-c(+) cloning/expression region

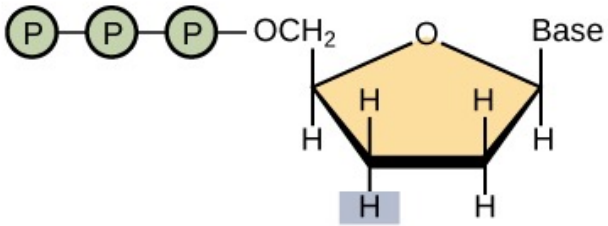
TP3: DNA Sequencing

ddNTPs

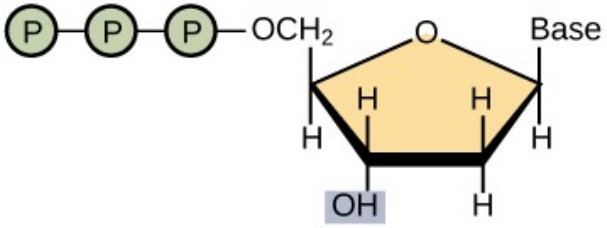
•Dideoxy, or **chain-terminating**, versions of all four nucleotides (ddATP, ddTTP, ddCTP, ddGTP), each labeled (either radioactive label or fluorescent label with a different color of dye)

Dideoxy nucleotides are similar to regular, or deoxy, nucleotides, but with one key difference**: they lack a hydroxyl group on the 3' carbon of the sugar ring**. In a regular nucleotide, the 3' hydroxyl group acts as a "hook," allowing a new nucleotide to be added to an existing chain.

Once a dideoxy nucleotide has been added to the chain, there is no hydroxyl available and no further nucleotides can be added. The chain ends with the dideoxy nucleotide, which is marked with a particular color of dye depending on the base (A, T, C or G) that it carries.



Dideoxynucleotide (ddNTP)

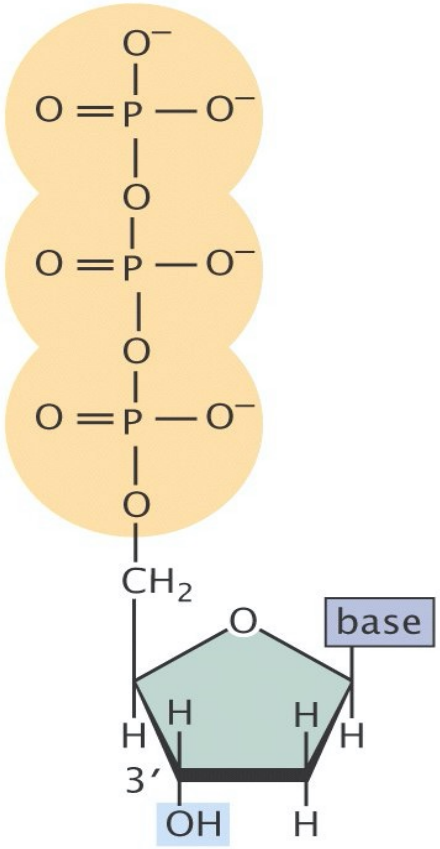


Deoxynucleotide (dNTP)

TP3: DNA Sequencing

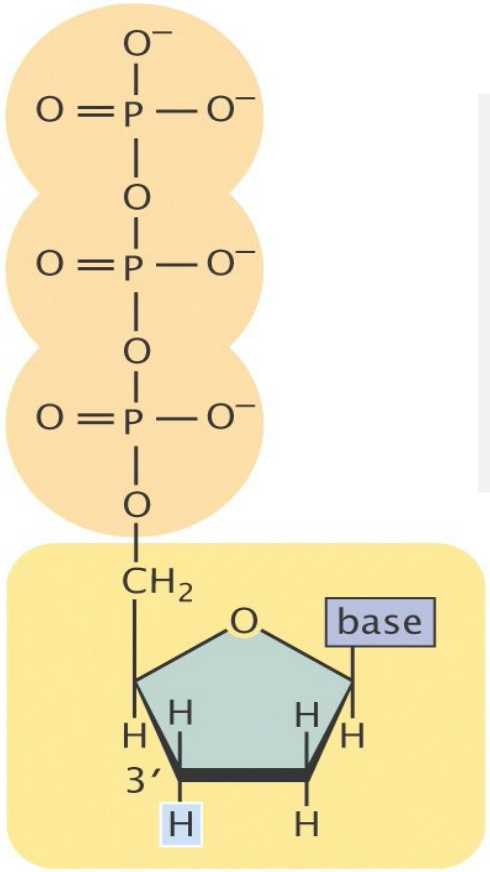
ddNTPs

A 3'-OH in normal DNA is necessary for elongation



Deoxyribonucleoside triphosphate (dNTP)

2'- deoxyribose



Dideoxyribonucleoside triphosphate (ddNTP)

2', 3'- dideoxyribose

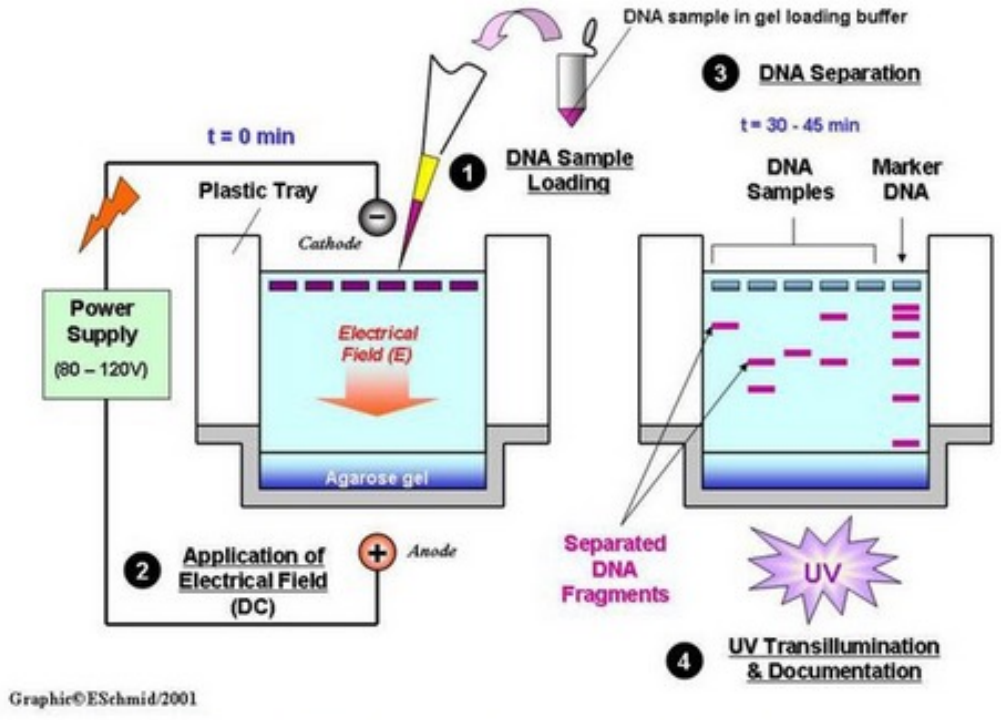
The dideoxy sequencing requires a special substrate for DNA synthesis

dNTP *vs* ddNTP

Didesoxirribonucleosido trifosfato (ddNTP)

TP3: DNA Sequencing

Manual DNA sequencing



Polyacrylamide gel electrophoresis separates ssDNA molecules that differ in length by just one nucleotide

Molecules are labelled with a radioactive protein or radioactive isotope, visualized by autoradiography producing a banding pattern

TP3: DNA Sequencing

Reading a sequencing gel

- You begin from the bottom where the smallest DNA fragments are,
- The sequence that you read will be in the 5'-3' direction,
- This sequence will complementary to the template DNA chain

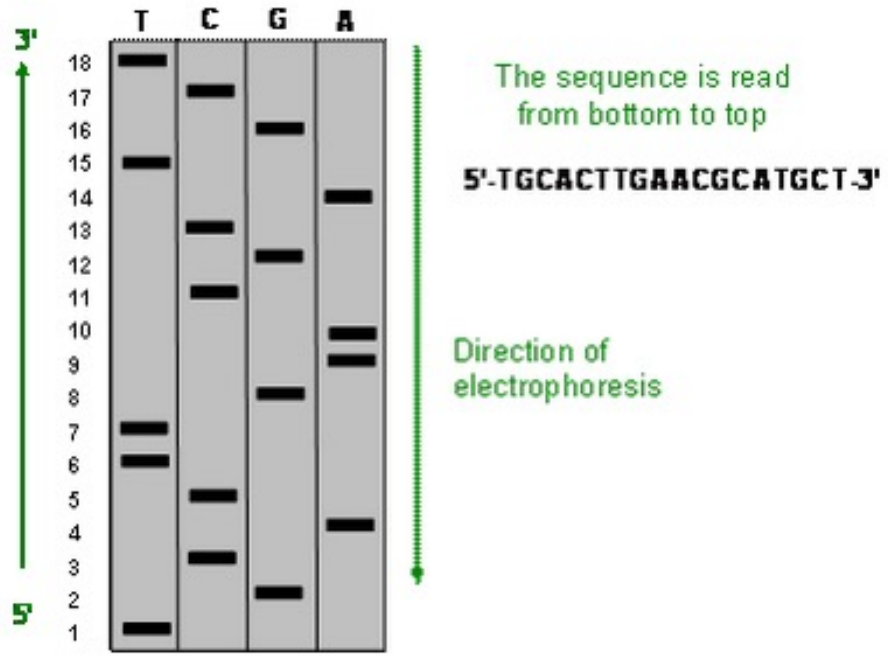
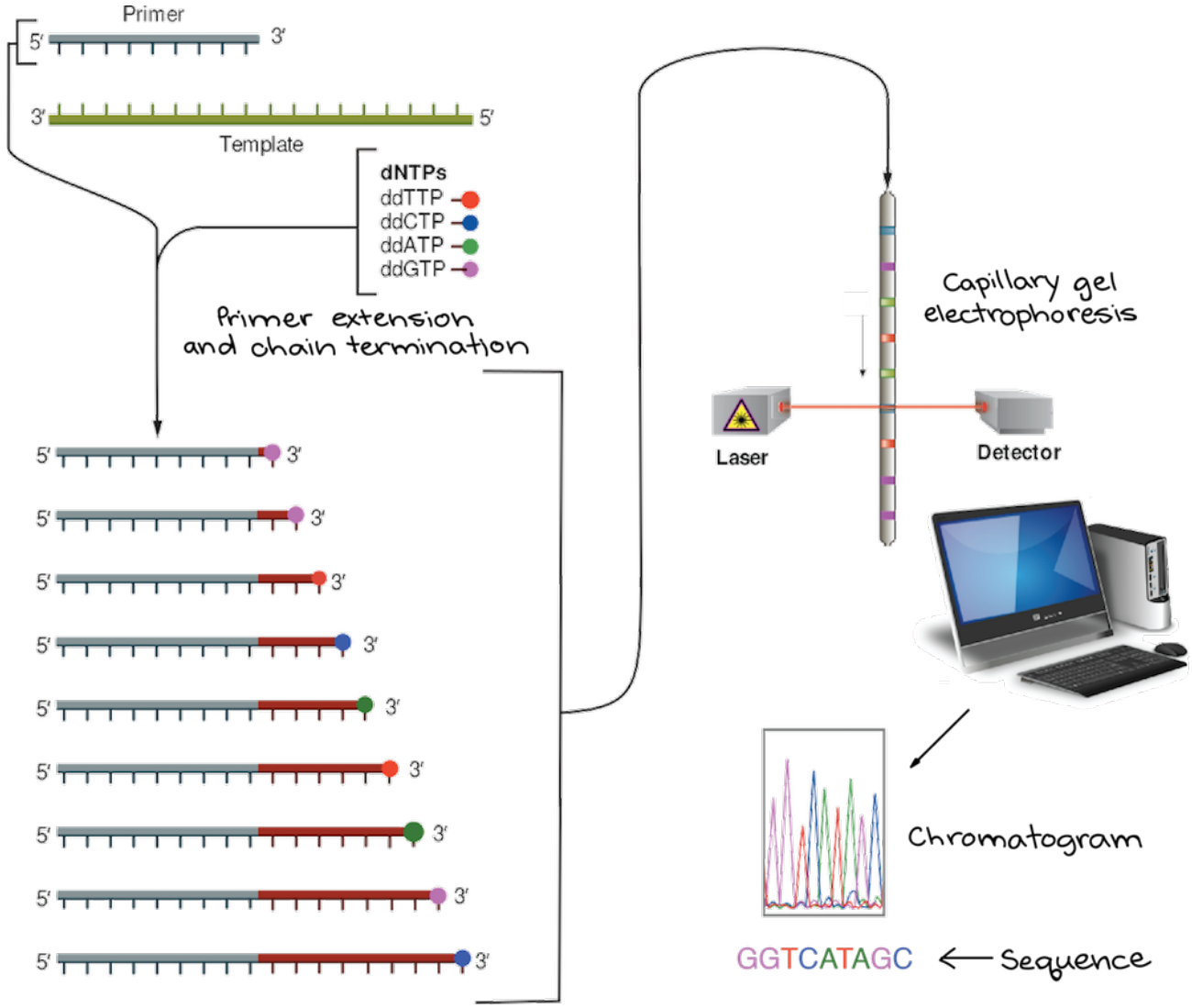


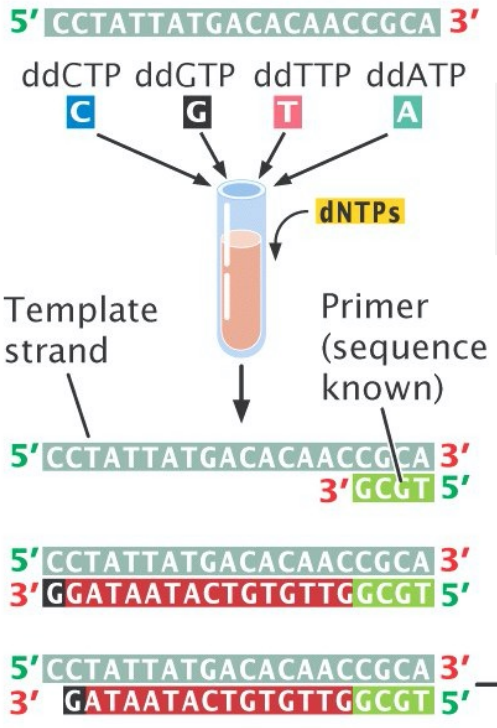
Figure 2. Reading the sequence from the gel.

TP3: DNA Sequencing

Automated DNA sequencing

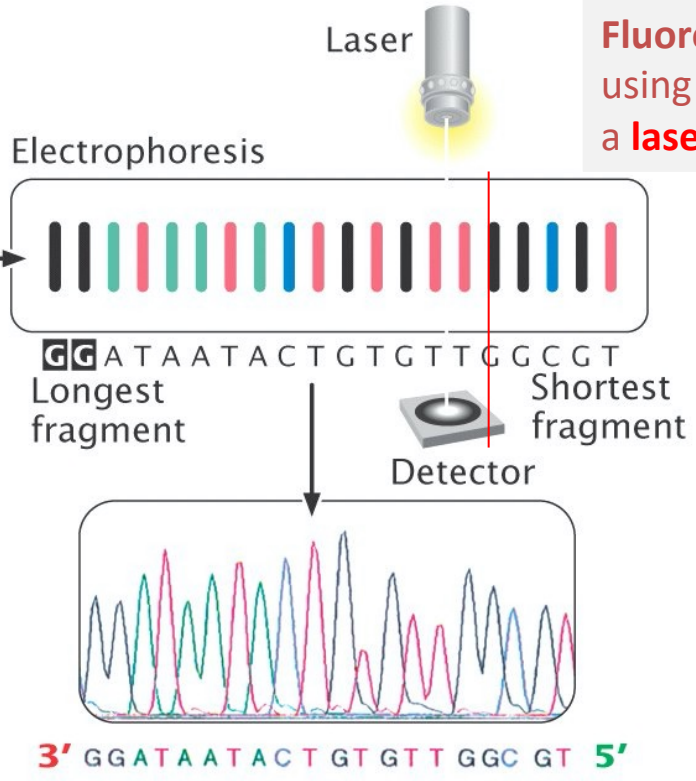


TP3: DNA Sequencing



Each of the four ddNTPs is tagged with a **fluorescent dye**

https://www.youtube.com/watch?v=e2G5zx-OJIw



Fluorescent dye detected by using a **laser** beam and a **detector**

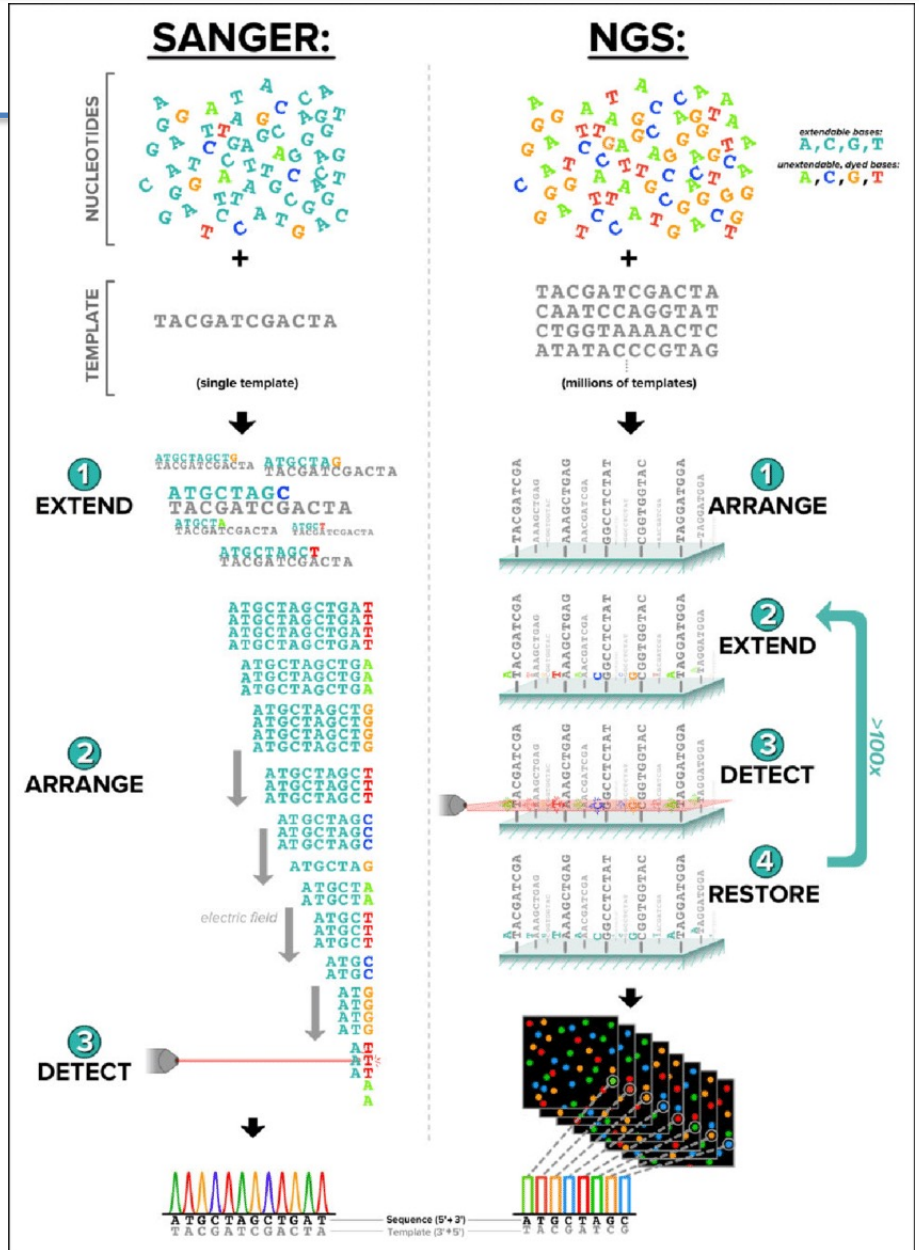
Denaturated DNA products are mixed and loaded into a **single well** on an electrophoresis gel.

The sequence information is directly **read** and **electronically stored** into the computer, which converts it into the complementary- target-sequence

TP3: DNA Sequencing

Sanger sequencing vs NGS

https://www.thermofisher.com/blog/behindthebench/when-do-i-use-sanger-sequencing-vs-ngs-seq-it-out-7/



TP3: DNA Sequencing

Sequencing technology advances

- 1868: Discovery of DNA
- 1953: Watson and Crick propose double helix structure
- 1977: Sanger sequencing
- 1985: PCR
- 2000: Working draft human genome announced (Sanger method)

$ human Genome

$3 billion

- 2005: 454 sequencer launch (pyrosequencing)
- 2006: Genome Analyzer launched (Solexa sequencing)
- 2007: SOLiD launched (ligation sequencing)

$2-3 million

- 2009: Whole human genome no longer merits Nature/Science paper

$250k

- 2011: Illumina sequencer (sequencing by synthesis)

$50k

- 2011: Ion torrent

$20k

$20k

$20k

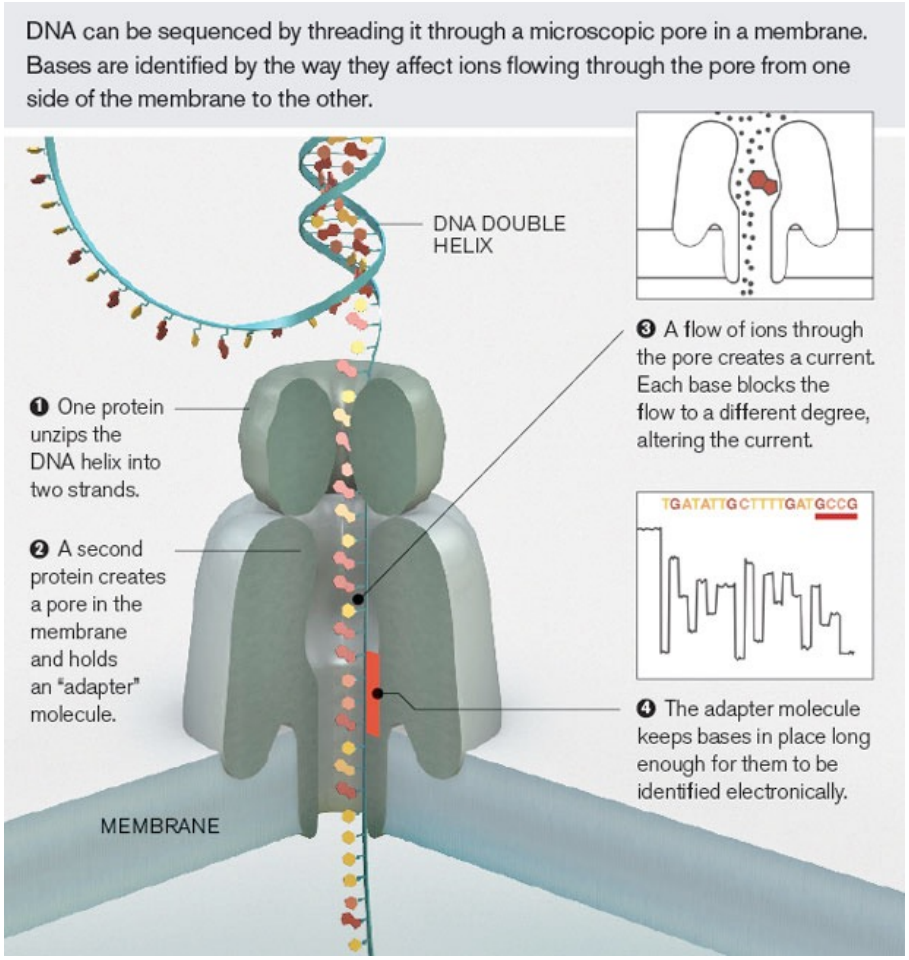
- 2011-18: 3rd generation sequencing: Pacbio, Oxford nanopore

?<$5k?

TP3: DNA Sequencing

NGS latest developments

Nanopore sequencing:

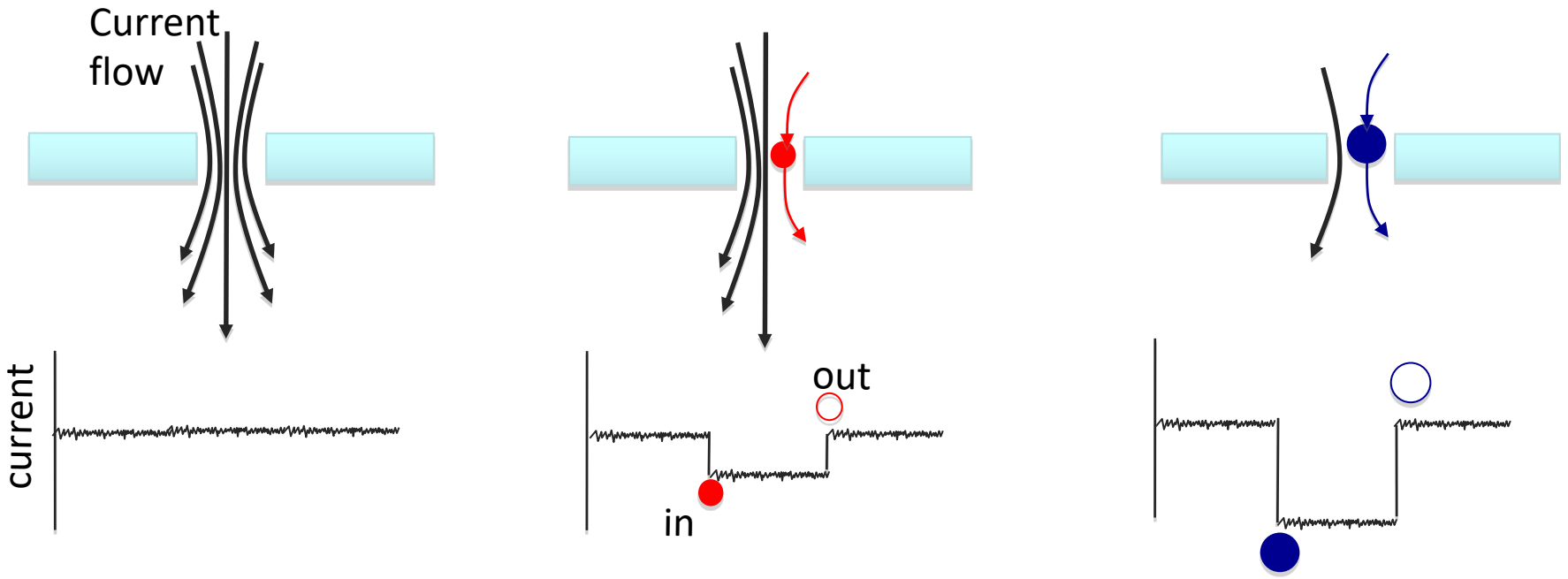


Determine the sequence of DNA fragments by passing DNA through a protein (or other) pore in a membrane

TP3: DNA Sequencing

NGS latest developments

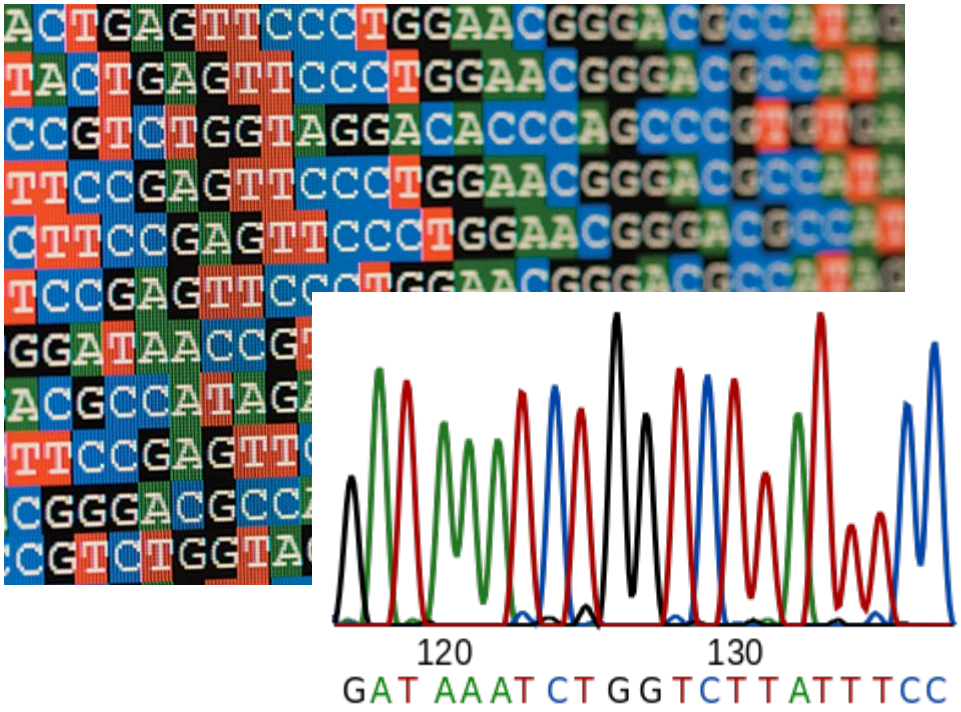
- Nanopore = ‘very small hole’
- Electrical current flows through the hole
- Introduce analyte of interest into the hole ➔ identify “analyte” by the disruption or block to the electrical current



https://nanoporetech.com/how-it-works

TP3: DNA Sequencing

Sequencing data analysis



?

Databases

Sequence alignment of different clones or reads

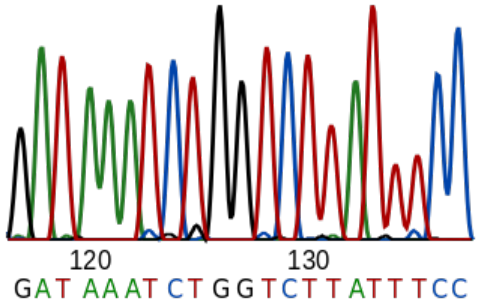
Nucleotide or peptide sequence comparison with other species (blast)

Sequence analysis for:

- Genome comparisons
- Restriction map
- ORFs
- Peptidic sequence
- Specific sequences (promoter, DNA-binding domains (ex. response elements), *stem-loop*, palindrom, direct and inverted repeats etc)
- % G/C
- *Codon usage* (*codon preference*)

TP3: DNA Sequencing

Sequencing data analysis



Major Sequence Repositories

GenBank or NCBI (all known nucleotide and protein sequences)
www.ncbi.nlm.nih.gov/Web/Genbank/
Ensembl (all known nucleotide and protein sequences)
www.ensembl.org/index.html

Genome Databases

Flybase (Drosophyla sequences and genomic information)
www.fruitfly.org
MGD (Mouse genetics and genomics)
www.informatics.jax.org
Grapevine
http://genomes.cribi.unipd.it/grape/
Arabidopsis
https://www.arabidopsis.org/

Genetic Maps

GBD (Human genes and genomic maps)
www.gbd.org
NCBI genome mapping
https://www.ncbi.nlm.nih.gov/probe/docs/applmapping/

Gene Identification and Structure

EID (Protein-coding, intron-containing genes)
mcb.harvard.edu/gilbert/EID/
Exint (Exon-intron structure of eukaryotic genes)
intron.bic.nus.edu.sg/exint/extint.html
TRRD (Regulatory regions of eukaryotic genes)
www.mgs.bionet.nsc.re/mgs/dbases/trrd4/

Protein interaction database

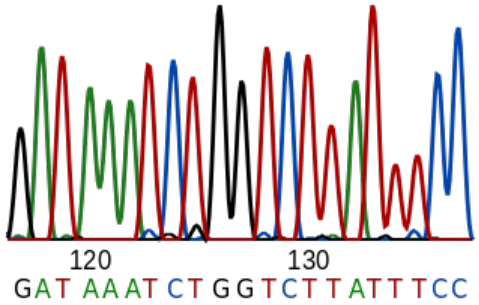
String
https://string-db.org/

Gene Expression

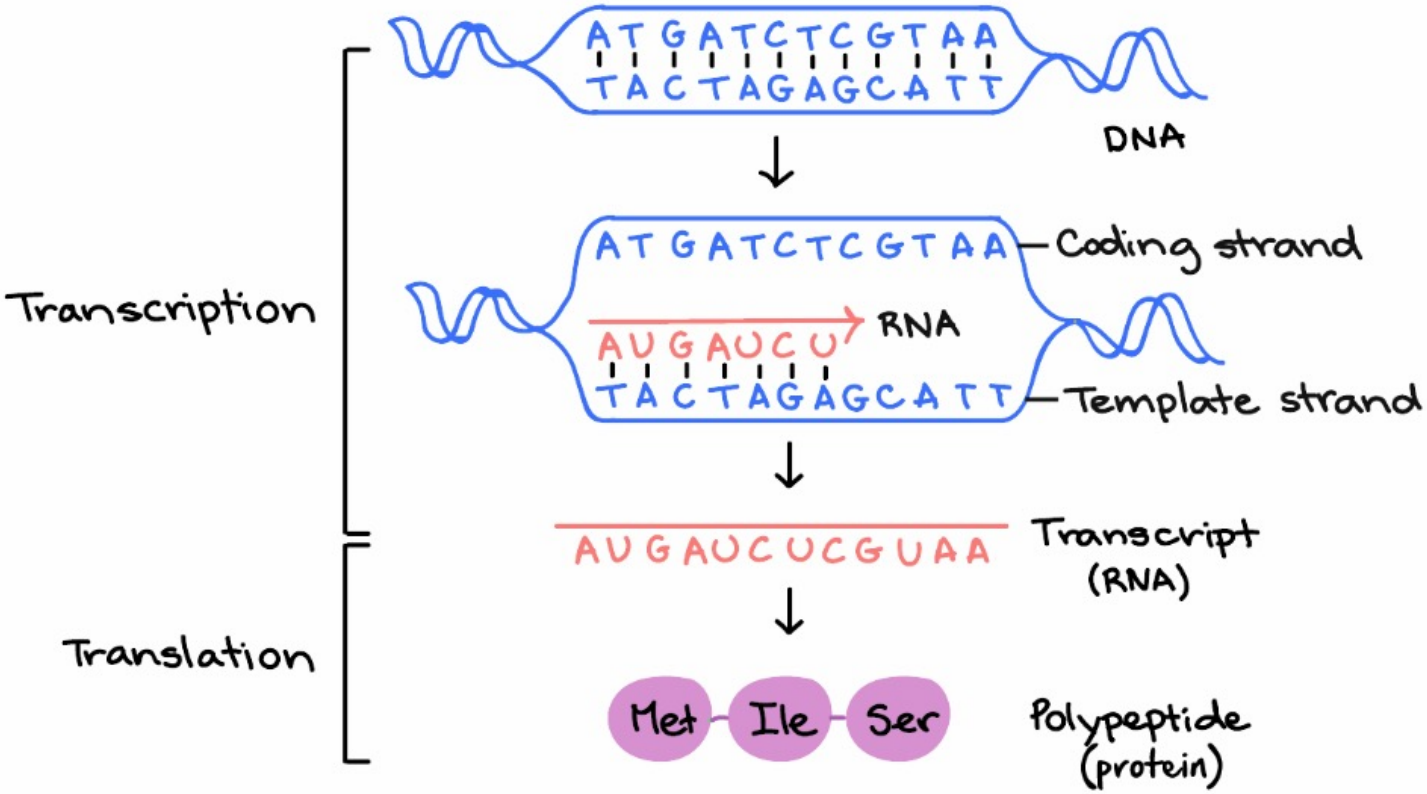
BodyMap (Human and mouse gene expression data)
bodymap.ims.u-tokyo.ac.jp
Tair
OPANDA

TP3: DNA Sequencing

Sequencing data analysis - ORF

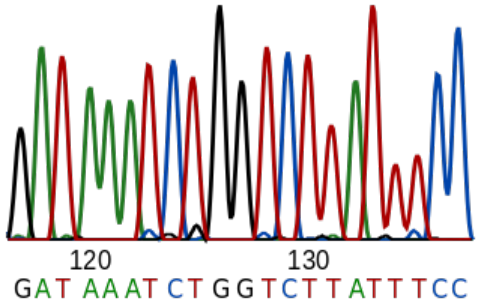


Definition of the **open reading frame**: (ORF) is the part of a reading frame that has the potential to code for a protein or peptide. An ORF is a continuous stretch of codons beginning with a start codon (usually **AUG**) and ending with a stop codon (usually **TAA**, **TAG** or **TGA**)



TP3: DNA Sequencing

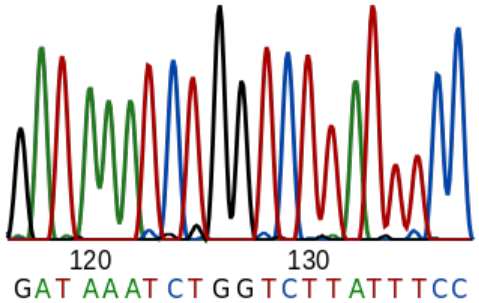
Sequencing data analysis - ORF

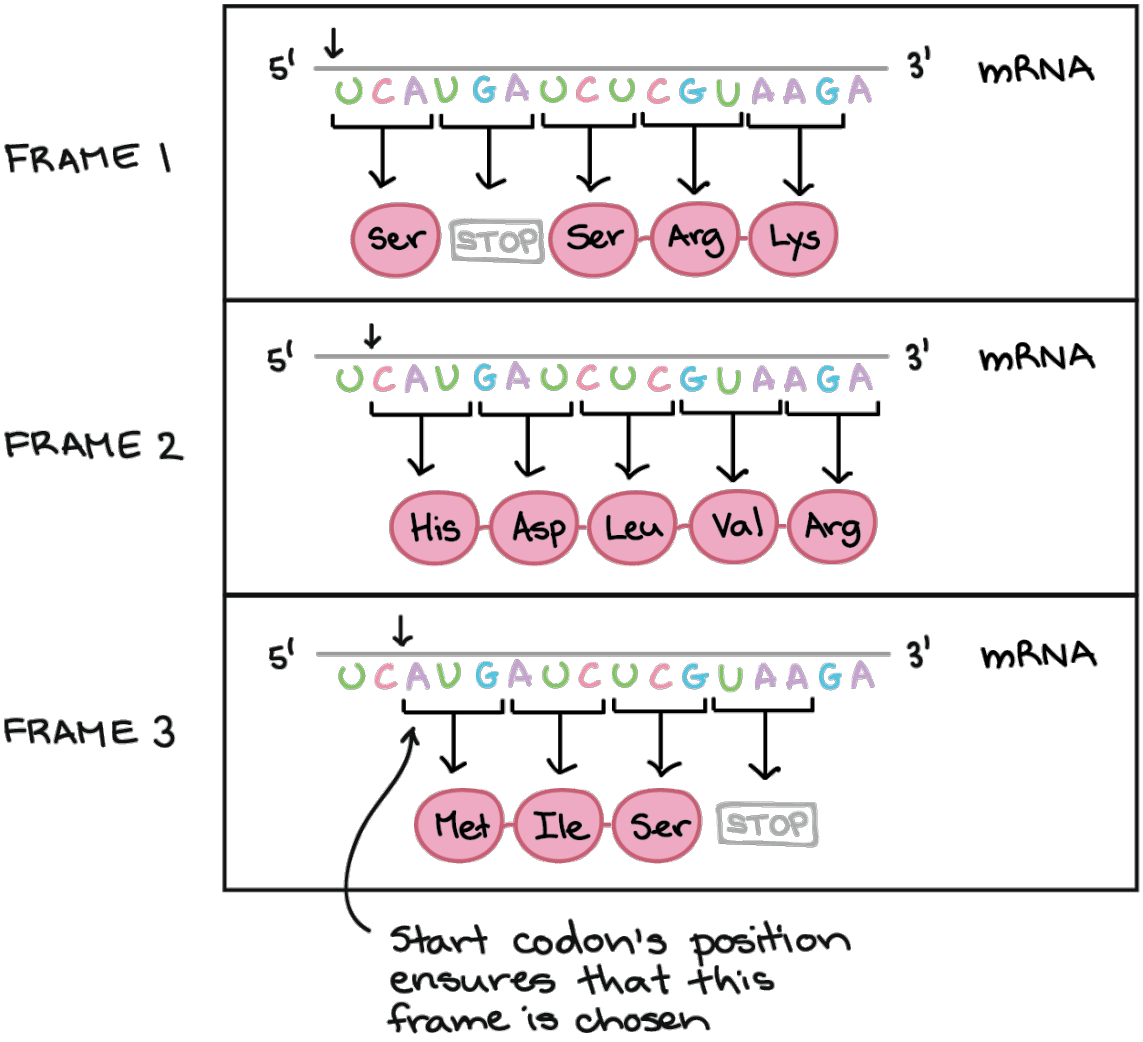


5'-Base		Middle	Base		3'-Base
	U(=T)	C	A	G	
U(=T)	Phe	Ser	Tyr	Cys	U(=T)
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Term	Term	A
	Leu	Ser	Term	Trp	G
C	Leu	Pro	His	Arg	U(=T)
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U(=T)
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala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

TP3: DNA Sequencing

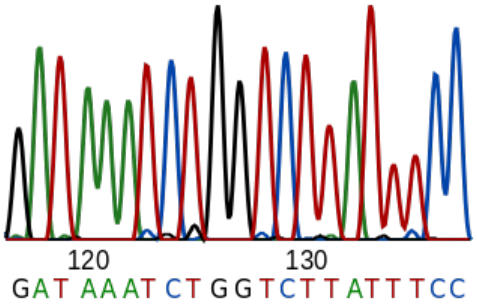
Sequencing data analysis - ORF

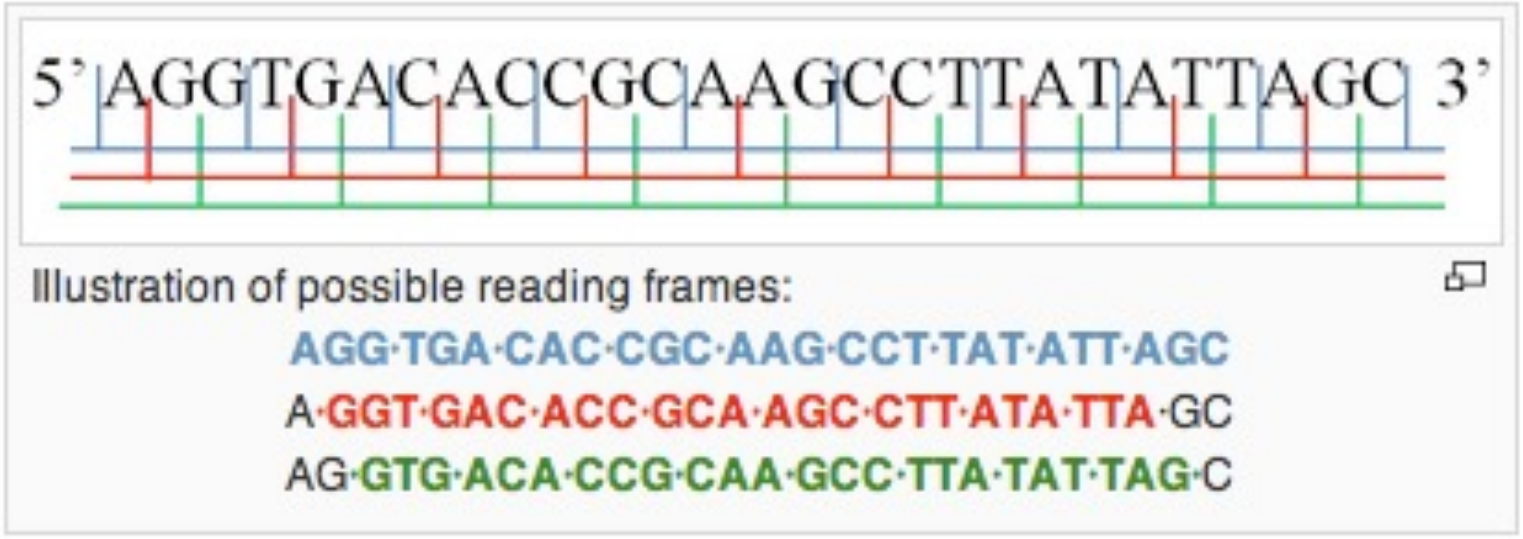




TP3: DNA Sequencing

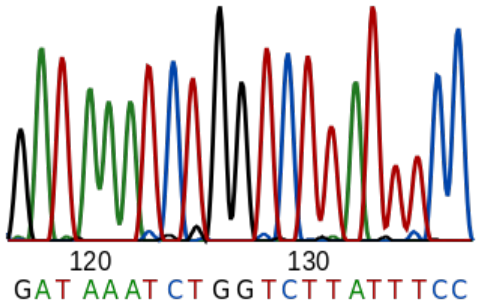
Sequencing data analysis - ORF





TP3: DNA Sequencing

Sequencing data analysis - ORF

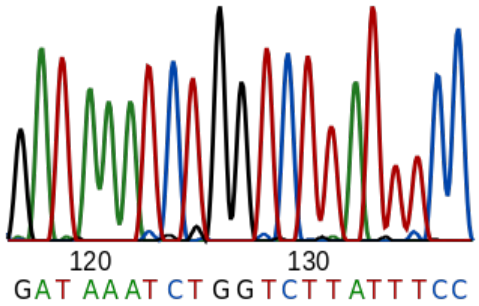



```
 5'                                                          3'
  atgcccaagctgaatagcgtagaggggttttcatcatttgaggacgatgtataa

1 atg ccc aag ctg aat agc gta gag ggg ttt tca tca ttt gag gac gat gta taa
   M   P   K   L   N   S   V   E   G   F   S   S   F   E   D   D   V   *
2  tgc cca agc tga ata gcg tag agg ggt ttt cat cat ttg agg acg atg tat
    C   P   S   *   I   A   *   R   G   F   H   H   L   R   T   M   Y
3   gcc caa gct gaa tag cgt aga ggg gtt ttc atc att tga gga cga tgt ata
     A   Q   A   E   *   R   R   G   V   F   I   I   *   G   R   C   I
```

TP3: DNA Sequencing

Sequencing data analysis - ORF



Possible Amino Acid Sequences (Forward)

Nucleotide Sequence

Possible Amino Acid Sequences (Reverse)

Gene 1

Gene 2

```
R S R A F W S P M S A A D S S * K A A P F T N R A S N R Q P R T A K
 D L G R S G R R C R R P T H L E R L H R S R T G R R T G N R G R R
 I S G V L V A D V G G R L I L K G C T V H E P G V E P A T A D G E
CGATCTCGGGCGTTCTGGTCGCCGATGTCGGCGGCCGACTCATCTTGAAAGGCTGCACCGTTCACGAACCGGGCGTCGAACCGGCAACCGCGGACGGCGA
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
GCTAGAGCCCGCAAGACCAGCGGCTACAGCCGCCGGCTGAGTAGAACTTTCCGACGTGGCAAGTGCTTGGCCCGCAGCTTGGCCGTTGGCGCCTGCCGCT
 R D R A N Q D G I D A A S E D Q F A A G N V F R A D F R C G R V A
  S R P R E P R R H R R G V * R S L S C R E R V P R R V P L R P R R
A I E P T R T A S T P P R S M K F P Q V T * S G P T S G A V A S P S

  S * S T K Q M W T T C R F P E R R C R * V A F V A S S G T V R G L
N R D R Q S K C G L H A D S L R G G V G K W L L S P R R E P F A G C
 I V I D K A N V D Y M Q I P * E A V S V S G F C R L V G N R S R V
AATCGTGATCGACAAAGCAAATGTGGACTACATGCAGATTCCCTGAGAGGCGGTGTCGGTAAGTGGCTTTTGTCGCCTCGTCGGGAACCGTTCGCGGGTT
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
TTAGCACTAGCTGTTTCGTTTACACCTGATGTACGTCTAAGGGACTCTCCGCCACAGCCATTCACCGAAAACAGCGGAGCAGCCCTTGGCAAGCGCCCAA
F D H D V F C I H V V H L N G S L R H R Y T A K T A E D P V T R P N
 F R S R C L L H P S C A S E R L P P T P L H S K D G R R S G N A P
  I T I S L A F T S * M C I G Q S A T D T L P K Q R R T P F R E R T

 S R P N I P A P Q R I A L S P W V S M E Y Y E I W R R Q P A V R R
  R A P T F Q R R N A S R C R R G C R W S I T R F G A V S P R C G A
V A P Q H S S A A T H R A V A V G V D G V L R D L A P S A R G A A R
GTCGCGCCCCAACATTCCAGCGCCGCAACGCATCGCGCTGTCGCCGTGGGTGTCGATGGAGTATTACGAGATTTGGCGCCGTCAGCCCGCGGTGCGGCGC
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
CAGCGCGGGGTTGTAAGGTCGCGGCGTTGCGTAGCGCGACAGCGGCACCCACAGCTACCTCATAATGCTCTAAACCGCGGCAGTCGGGCGCCACGCCGCG
  D R G L M G A G C R M A S D G H T D I S Y * S I Q R R * G A T R R
Q R A G V N W R R L A D R Q R R P H R H L I V L N P A T L G R H P A
 T A G W C E L A A V C R A T A T P T S P T N R S K A G D A R P A A

A V S F L A R N I A Q L G L H L F E R K D D A D R K R L T D H P L A
 R C R S W R A T S R N S V C T C S S A R T M P T A S G * P T T R S
  G V V P G A Q H R A T R S A P V R A Q G R C R P Q A V D R P P A R
GCGGTGTCGTTCCTGGCGCGCAACATCGCGCAACTCGGTCTGCACCTGTTCGAGCGCAAGGACGATGCCGACCGCAAGCGGTTGACCGACCACCCGCTCG
....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|....:....|
CGCCACAGCAAGGACCGCGCGTTGTAGCGCGTTGAGCCAGACGTGGACAAGCTCGCGTTCCTGCTACGGCTGGCGTTCGCCAACTGGCTGGTGGGCGAGC
 A T D N R A R L M A C S P R C R N S R L S S A S R L R N V S W G S
  R H R E Q R A V D R L E T Q V Q E L A L V I G V A L P Q G V V R E
R P T T G P A C C R A V R D A G T R A C P R H R G C A T S R G G A R
```