# Bioinformatics and Systems Biology (BSB)
# Introduction - Molecular Biology (4h) and Biochemistry (4h)

TEACHERS: **Laura Marchetti, Eleonora Da Pozzo**

E-mail: laura.marchetti@unipi.it, eleonora.dapozzo@unipi.it

In-person or online appointment: **upon request by e-mail**

*N.B. These slides contain figures taken from textbooks for educational use only. Their use is covered by copyright. Therefore, any type of reproduction and dissemination is prohibited.*

REFERENCE BOOKS:

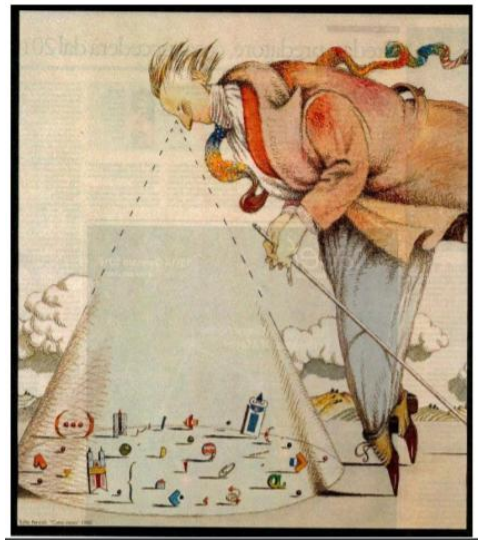
- **Molecular Biotechnologies, Terry A. Brown, Zanichelli, 2° edition (2019)**
- **Molecular Biology of the gene, Watson-Baker-Bell-Gann-Levine-Losick; English Version by Pearson Education, 7° edition (2014)**
- **Wilson and Walker's Principles and Techniques of Biochemistry and Molecular Biology, Cambridge University Press (2018)**
- **Lehninger Principles of Biochemistry, David L. Nelson, Michael M. Cox, Aaron A. Hoskins, W. H. Freeman & Company, 8° edition**
- **Slides available**
- **Reviews/articles from the literature (DOI is given in the slides)**

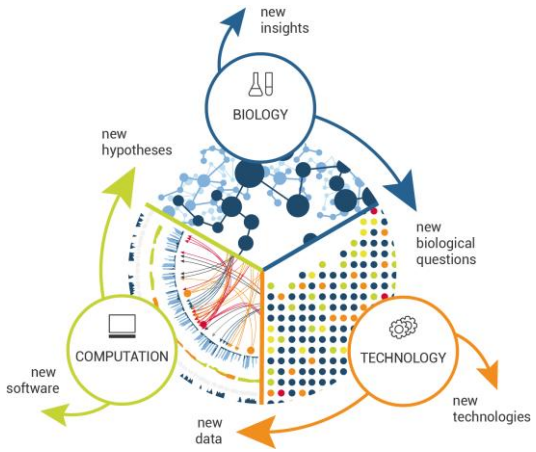
EXAM:

**Written test possibly «in itinere», i.e. one week after the end of the 8 hours: Friday 3rd October, at 14:00**

# Some important definitions…

**SYSTEMS BIOLOGY:** this name refers to a **holistic approach** to decipher the complexity of biological systems. It assumes that the networks that form the whole of living organisms are more than the sum of their parts.
It is a collaborative and **multidisciplinary** subject, integrating many scientific competences – biology, computer science, engineering, bioinformatics, physics and many more.
It is based on the possibility to **experimentally reduce the complexity of a biological system** (e.g. by molecule fragmentation) and **convert the biological complexity into a complex computational problem**.






*System biology work cycle*

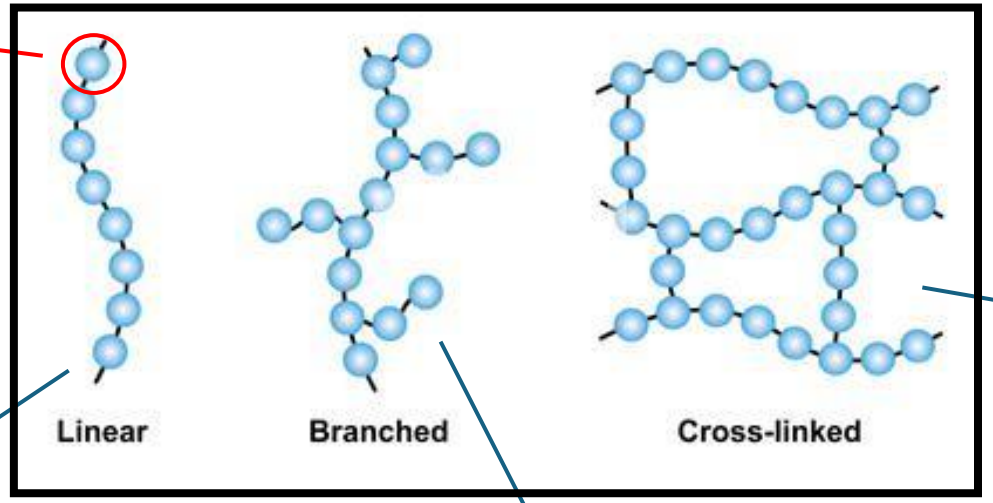
**OMICS:** The term "omics" is derived from the Latin suffix "ome" meaning **mass** or **many**. Thus "omics" studies involve a mass (large number) of data rather than one or a few. In biomedicine, this translates in the study of the bulk of biomolecules (either DNA, RNA, protein, lipid, metabolite, etc) contained in a biological system (either a cell, tissue, organ, organism).

**PRIMARY SEQUENCE:** omics aim at the identification of the primary sequences of biomolecules, i.e. the exact specification of their atomic composition. For typical **unbranched, uncrosslinked biopolymers (DNA, RNA, protein),** this is equivalent to the ordered string of their monomeric subunits (nucleotides, amino acids).

DOI: 10.1016/j.trac.2006.10.007; 10.1016/j.cell.2005.04.020

# Different types of polymers in nature…

Monomer or building block




Some specialized forms of proteins like collagen

DNA
RNA
Protein

Sugars

For linear polymers, all the information for the biomolecule build-up resides in the ordered assembly of the building blocks. The assembly is covalent (i.e. they are connected by a type of chemical bond formed when two atoms share one or more pairs of electrons).

Linear polymers differ for their combination of building blocks and length.

Such combinations give rise to a plethora of different 3D structures (folds/folding) and functions.

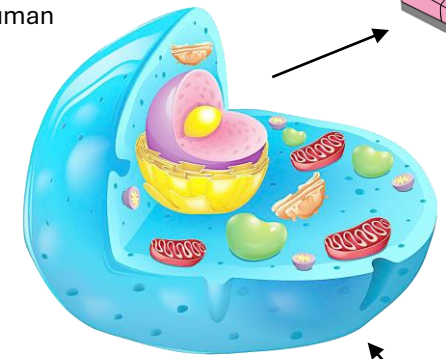
# Biomolecule numerosity

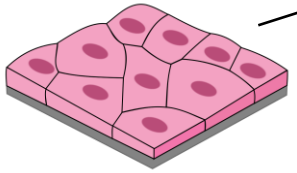
**ORGANS**
78 in human

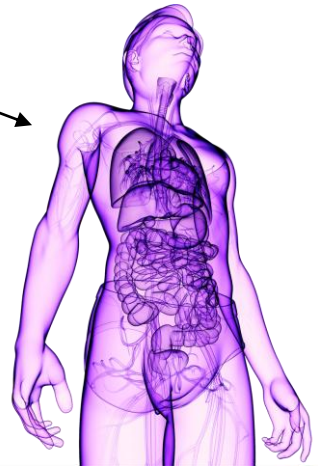
**LIVING HUMAN ORGANISM**

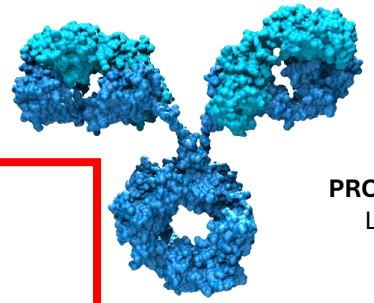
**TISSUES**
4 types in human

**CELL**
~$4x10^{13}$ cells in human



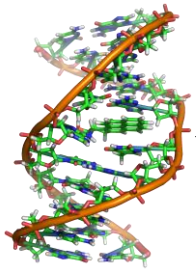


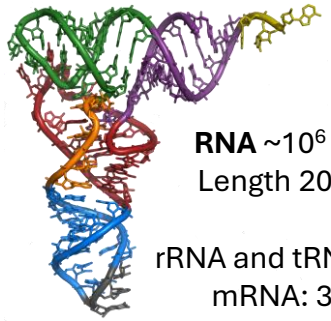




**PROTEINS** >$10^5$ molecules
Length 50-35,000 aa

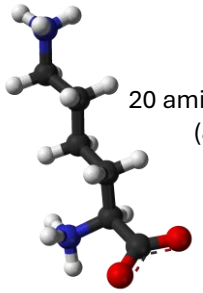
**DNA** 46 paired molecules = CHROMOSOMES
Organized in ~ $3*10^4$ genes
Length $50x10^6$-$130x10^6$ nt
$6x10^9$ paired nucleotides = 2 meters of length!



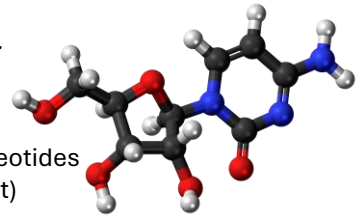


**RNA** ~$10^6$ molecules
Length 20-20,000 nt

rRNA and tRNA: 80% ($10^6$)
mRNA: 3-5% ($10^5$)
miRNA/ncRNA: $10^3$-$10^4$



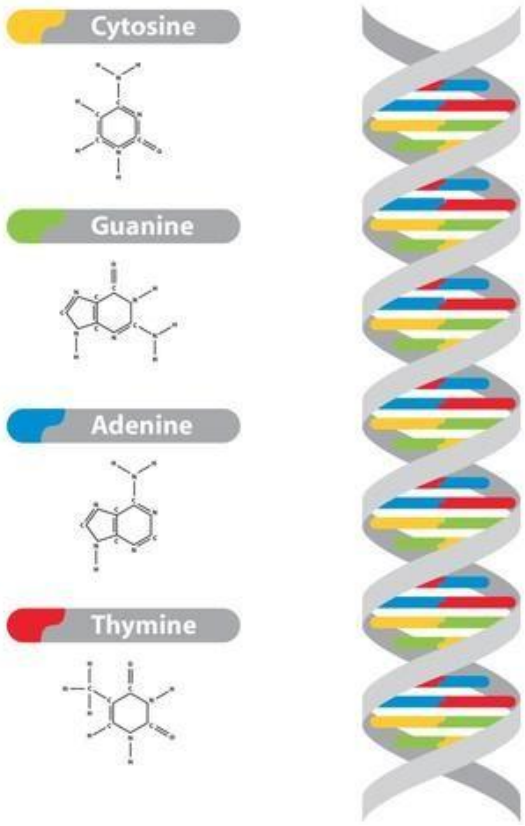
20 amino acids
(aa)



4 nucleotides
(nt)

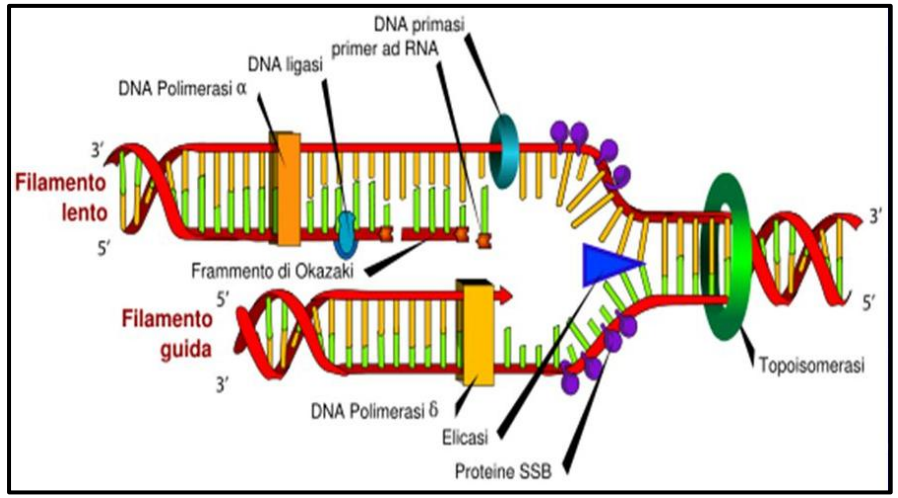
https://bionumbers.hms.harvard.edu/search.aspx

# DNA



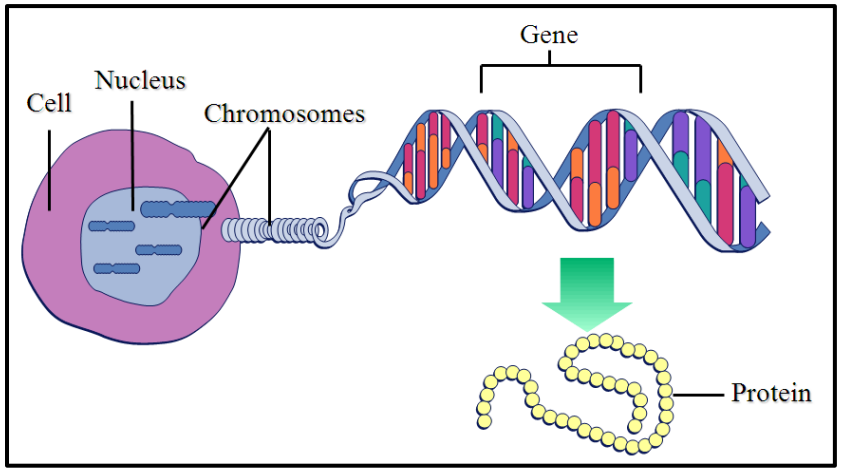

Twisted ladder: the **sugar-phosphate** backbones form the **rails**, while the **base pairs** form the **rungs**

## DNA is replicated



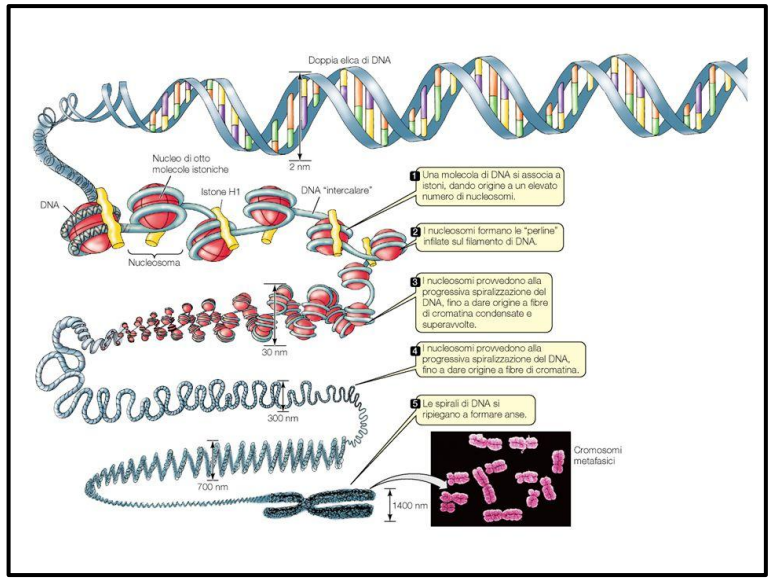

DNA is replicated easily, with a mechanism exploting each strand as a template and this mechanism has been converted in a technology for sequencing

## DNA contains genes



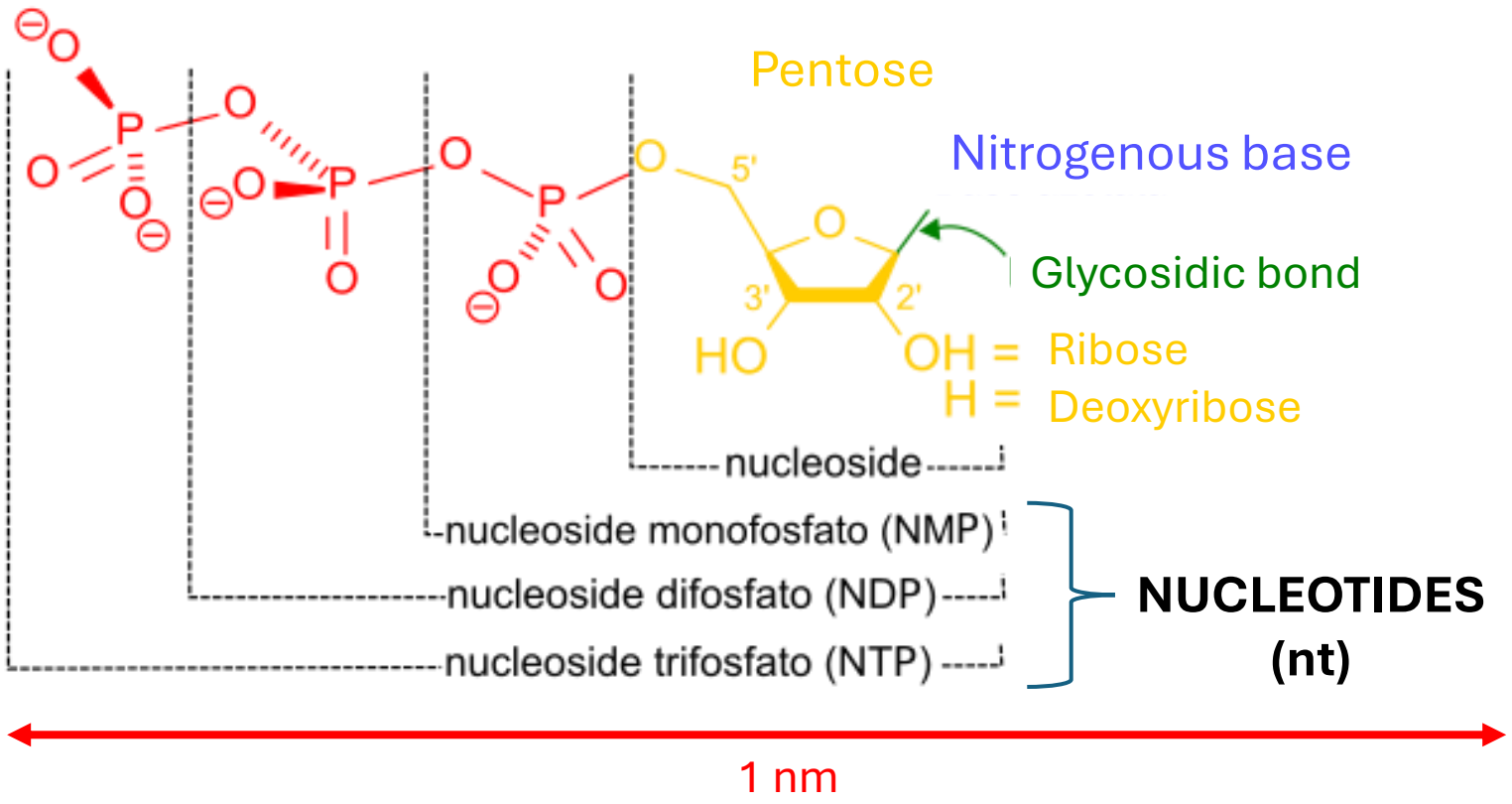

DNA is the molecule that stores genetic information. Its functional units are called genes, and each gene contains the instructions needed to produce a protein.

## DNA is compacted

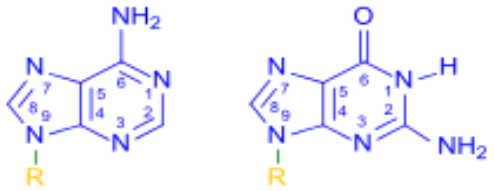



DNA is compacted thanks to intrinsic modifications and specialized proteins; the structured form of DNA+protein is named chromatin and its compaction grade determines whether or not a gene is expressed, i.e. it is converted into a protein

# DNA/RNA building blocks

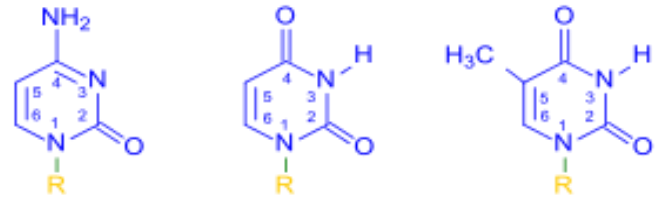



Purin bases



Adenina Guanina

Pyrimidine bases

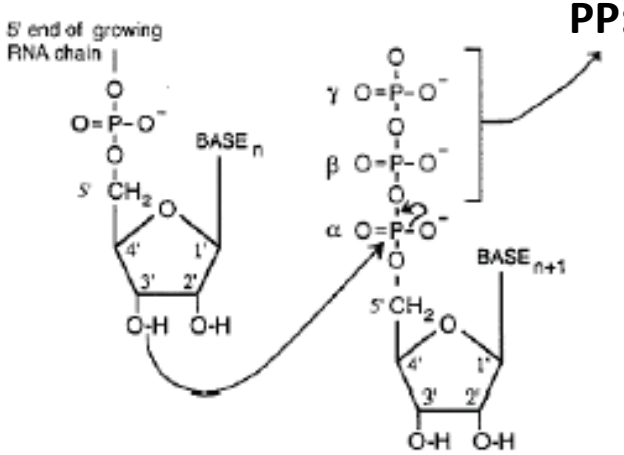


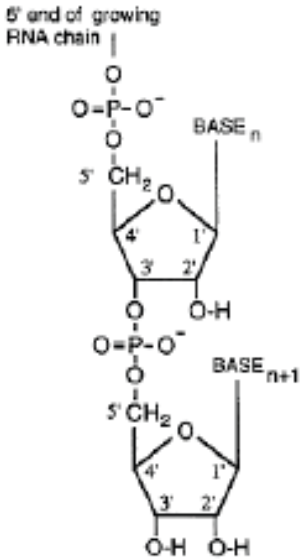
Citosina Uracile Timina

- cyclic compounds containing Nitrogen (N)
- **delocalized π-electrons**, which can interact with each other
- These interactions lead to **hydrophobic base stacking**, a key stabilizing force in the DNA double helix

# Nucleotide polymerization

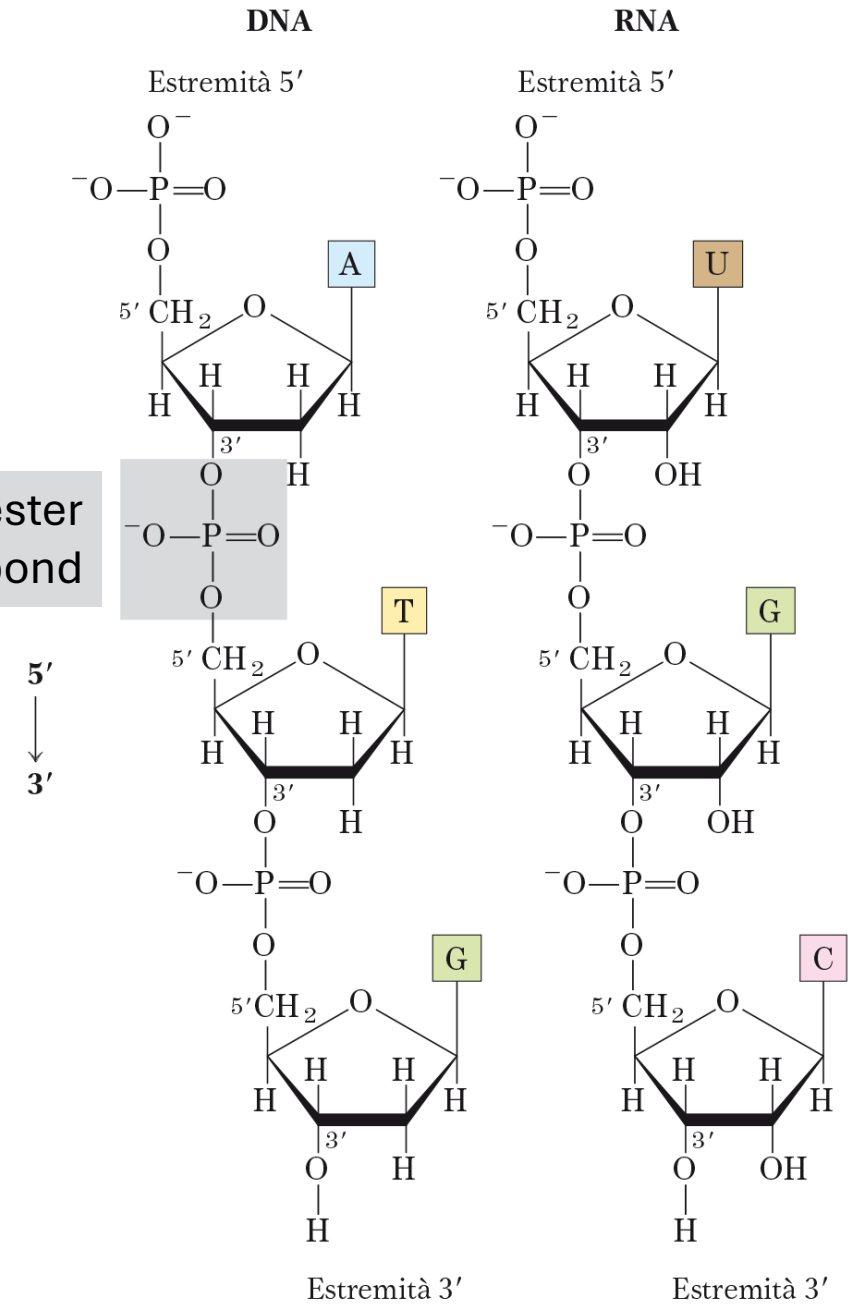
PP: pyrophosphate



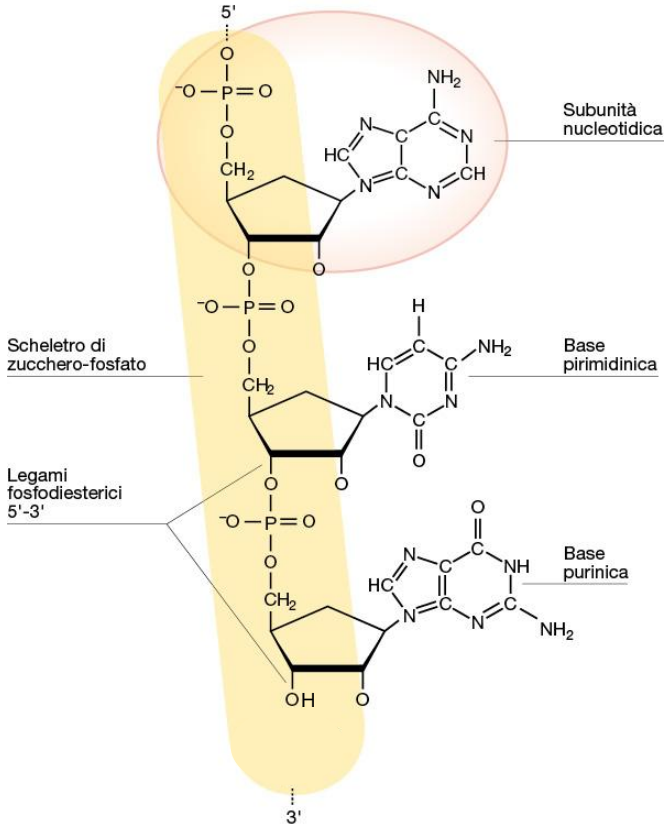




(d) polymerization reaction

Phosphodiester bond




Oligonucleotide: <50 nucleotides
Polynucleotide: > 50 nucleotides



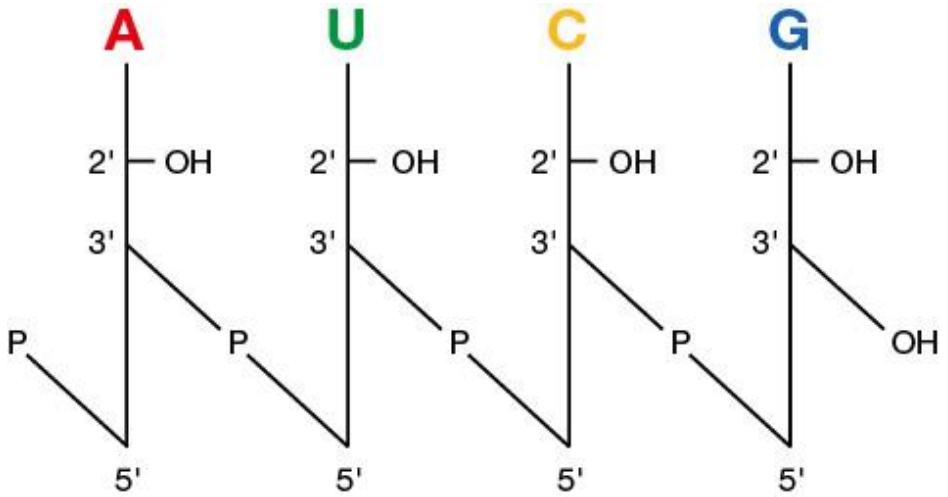




**Figure. A polynucleotide chain has a repetitive structure. It consists of a backbone in which sugars and phosphate groups alternate, linked together by 5' → 3' phosphodiester bonds. From this backbone, the nitrogenous bases protrude laterally.**

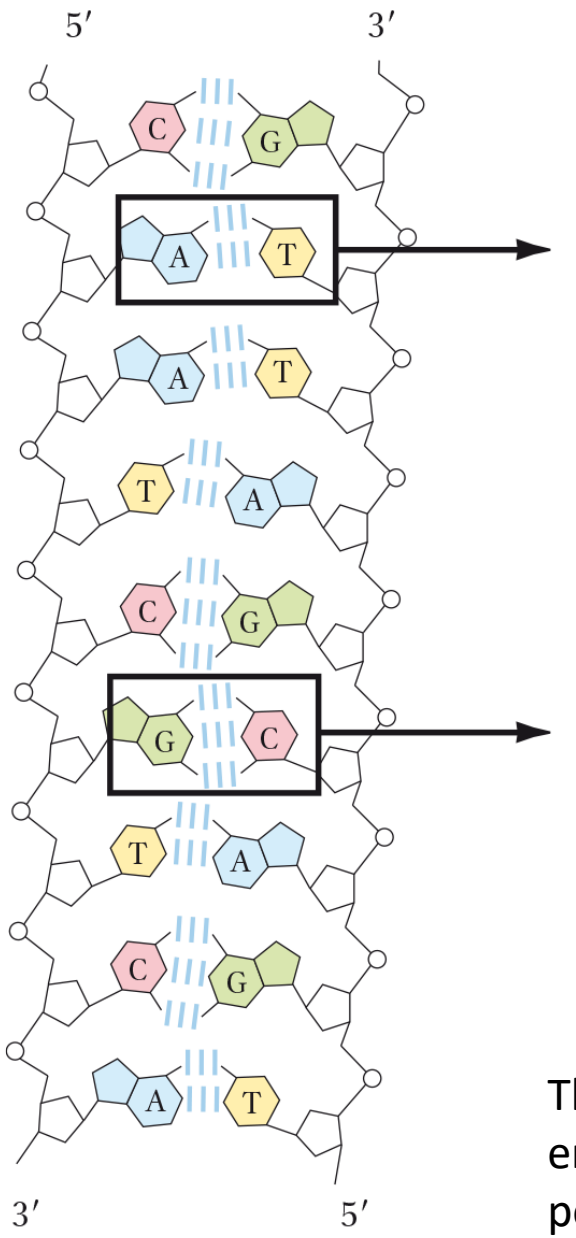
The sugar-phosphate backbone is hydrophilic: it interacts well with water because they can form favorable interactions (like hydrogen bonds) with it, due to the charge. Also termed “polar”

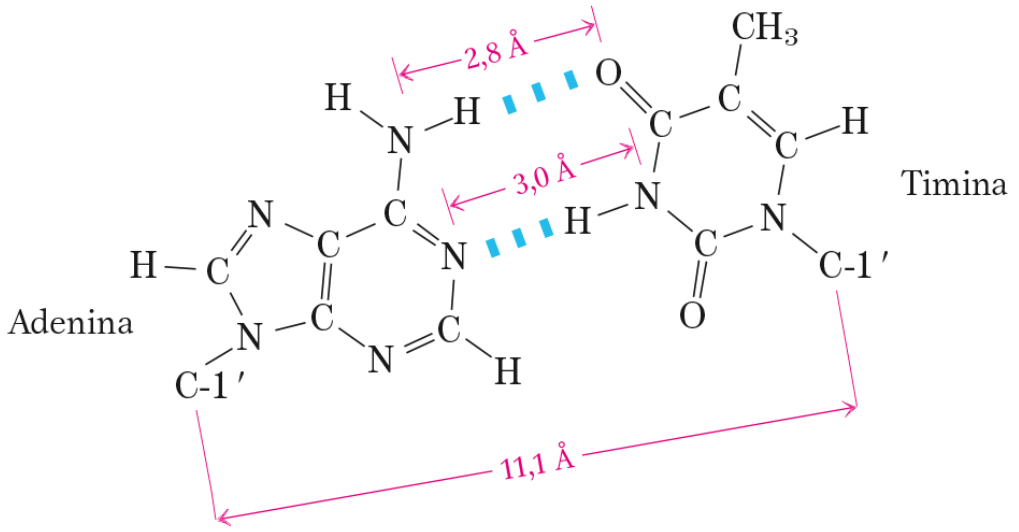
The nitrogenous bases are hydrophobic: do not interact well with water. Instead, they tend to cluster together to avoid water. They are referred to as “non-polar”

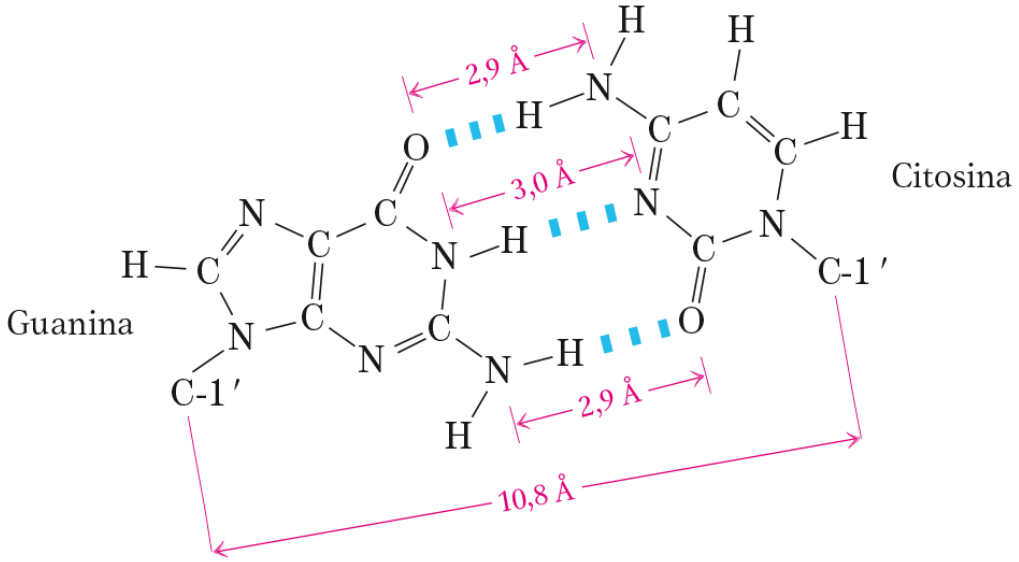
5’ end: typically contains a free phosphate
3’ ned: typically contains a -OH

The polynucleotide chain has a polarity 5’ -> 3’

# DNA Secondary structure




The formation of **hydrogen bonds between nitrogenous bases** ensures the maintenance of a constant distance between polynucleotide chains → leads to **DNA double helix**

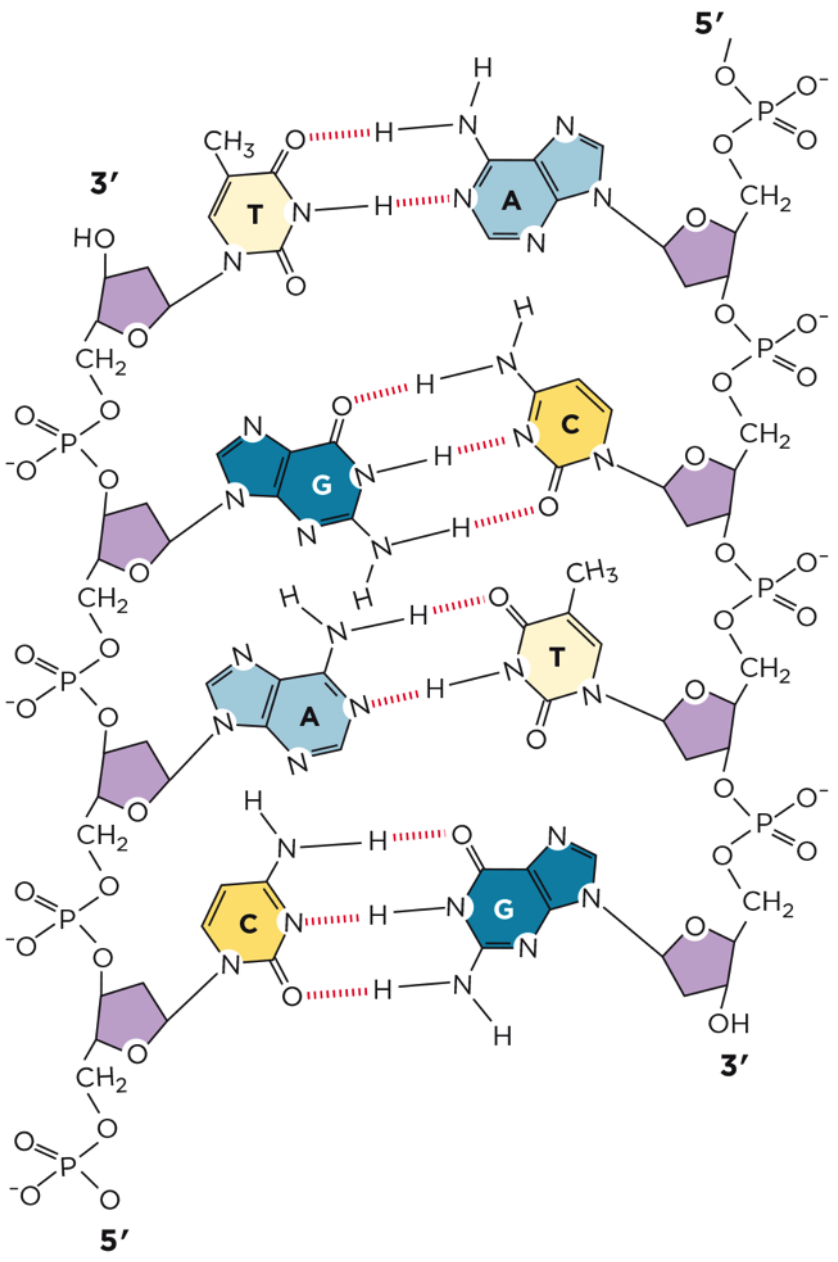
# Antiparallel and complementary strands

- The two DNA strands run in **opposite directions** (5′ → 3′ and 3′ → 5′): they are **antiparallel**.

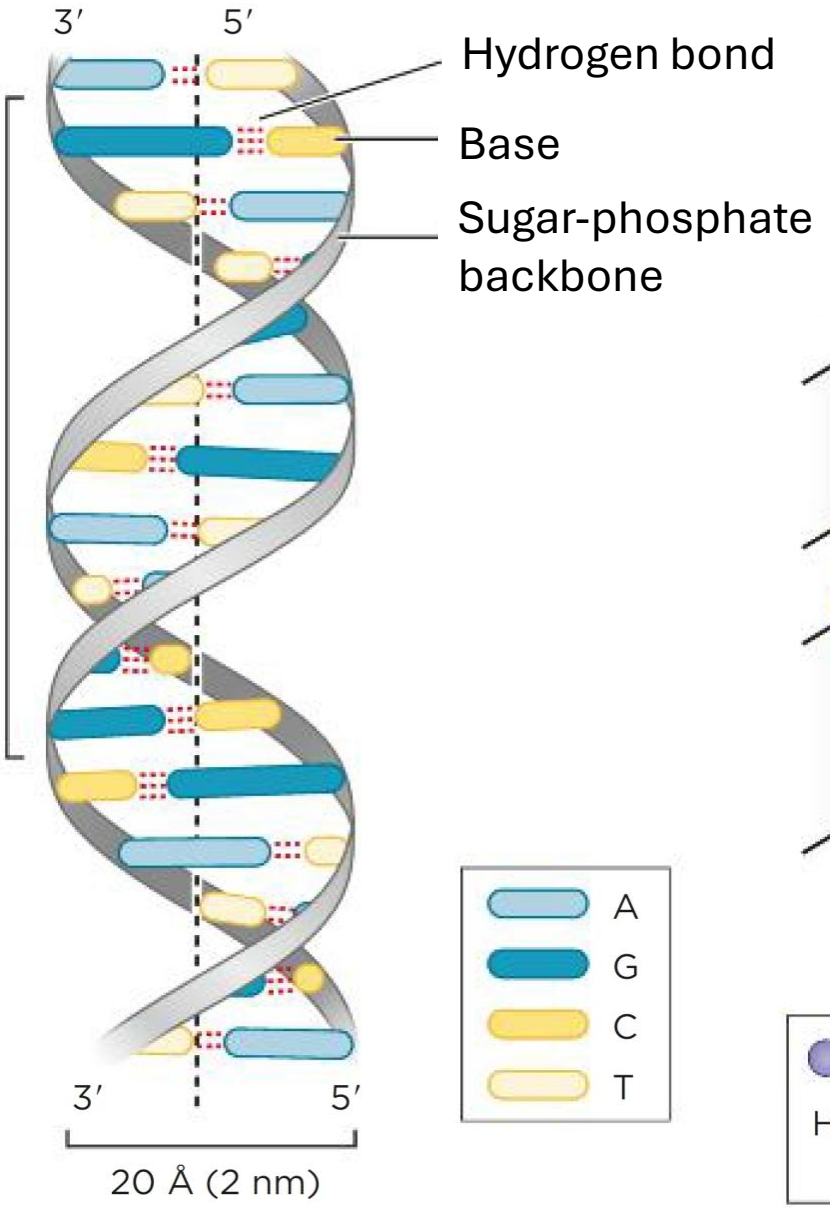
- Base pairing is **complementary**: adenine pairs with thymine, guanine pairs with cytosine. Base pairing is also referred to **Hybridization** and is a distinguishing feature of nucleic acids.

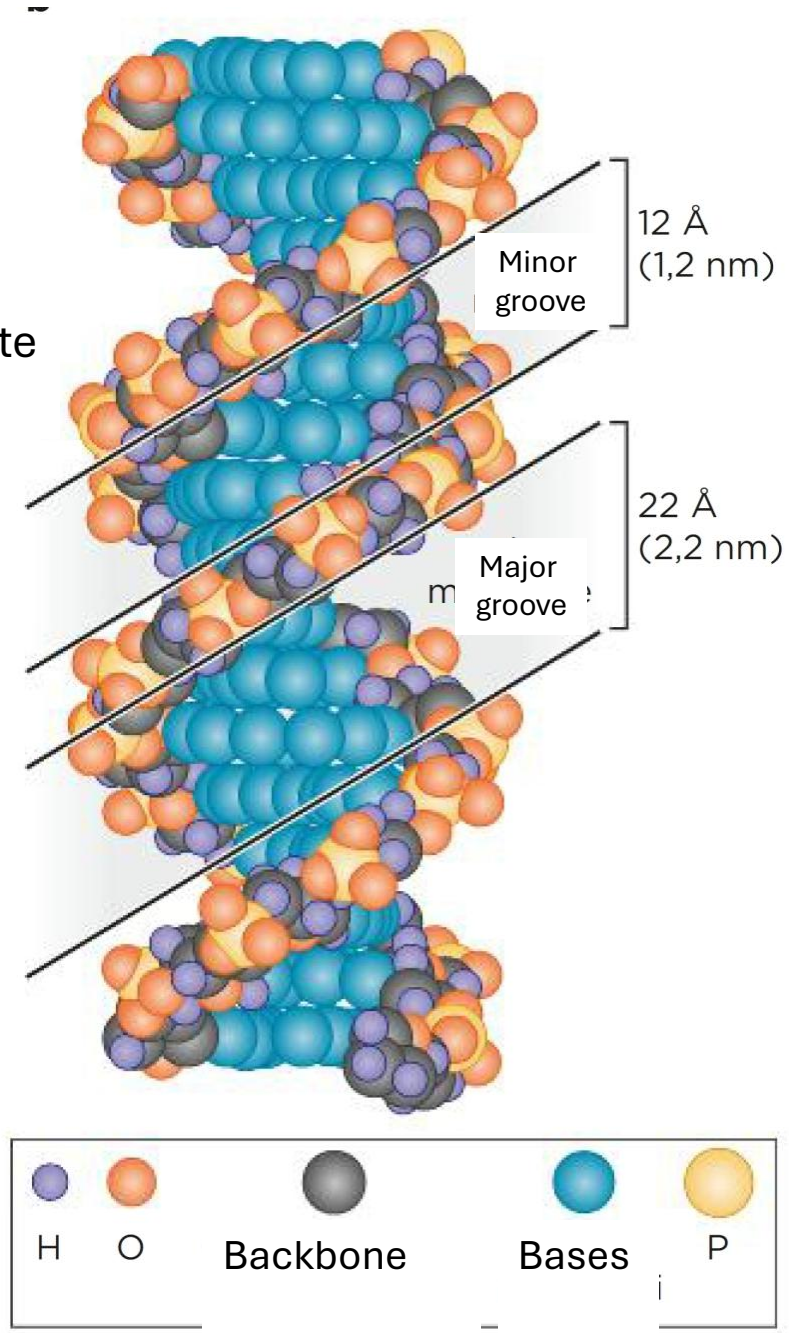
- This arrangement ensures **stability** and **accurate replication** of the genetic material.

- The content of the two strands is **redundant**: once you know the sequence of one strand, you will always know the complementary one, too.



# DNA Tertiary structure



# All the information is in the DNA primary sequence

The FASTA format is a text-based format for representing DNA sequences, in **which nucleotides are represented using single-letter codes corresponding to the nitrogenous bases**. Used in all DNA databases.

NCBI Reference Sequence (Nucleotide database)

```
>NC_000006.12:151654148-152129619 Homo sapiens chromosome 6, GRCh38.p13 Primary
Assembly
TATTGATTTTTGTGTAAACATGTGTTTGTATATATCTATAACGAGAACTCAAGTCATACTGTAATCCTAT
TTTGTAAACTGACTTTTTTCCTTTATCAGTATATCAAGATTATTTTCCCACATCATTTGACATTTTTTCT
ACAGTGTAATTTAATGGCTACATTGTTTTCTATCCTATGAATATATCAAACCTATTTCTTAAAAACCCTA
CTCAGGGATTTTAAAAAATAAAAACGATGTTTTAATATTATAAAGATTCAGTGAGGTATATTCTTATACG
TACACATTTCTAAGGTTTGAGTTCTTACAAGATGCTGAACTAGCTAAGACTACTGGTTCTCATCTGTCAC
ATAGGGAAAAATTATAGAAGGAAAACATCAAGATTTGGAAAAATCTGTGAGAATTGTTTTGCATTAGTGT
GTAGGTGTGTGTGTTGGGGTGGTGGCTGCAGCTTGGGGCAGAGGCCTCAGGTGTGGCTGTGGAGTGATCA
GATAGAGTTTTTGGAGTTCGGCTTTTGCCCCAGGACACTTGGTGCCTGCCCCCAGAGCTGCAGCCCAGAA
GGCCGTTCTCAGAGGTGAAGTCCAGGCAGTGAGGAGCTGTCTGCCAGTAGGCAGTTGAAGAAAAAAAATG
AGCTAGAGGAAAAAAACAAAAAAACAAAATCTCCTTCTAATGCTGCCAGGCTGCCGGGAGCTGGAAATGA
AGCACTGACAGGAGTGGGTATTTCATGGTGAAGGGAATAATCAACTGGTTTTTTTGGTACCCAAGACTTT
CCACCTTCACACACACACATGAGATGCTTTGAAATAAAGATAGTCACTTGACTTAGTAAAGTTTGTTGAC
ATAAAAATATGAGAAATACCAAAGAATACAAAAAGGAAAACTTCGTTAATATTATTCAGACTTAAAATTC
CAGATTGTATCAACATTAAGGGGGTTGATGAAAACATGGGAGAAAGCCAAGGGACGTGAGATCGGGCTCA
ATTCTTGACTTGCTGGGGGAAGGTATCAACACAGAACTTTTAAGAATTAGAAGGCATTAAAAAGAAATAG
AAATCCTGAATCAAATTGAAACAGTAAAATAAAATAGTCCAAAGATGTGTAAATATATCACTATCACAAT
```

HEADER

SEQUENCE

https://www.ncbi.nlm.nih.gov/

```
5' ATGACGACTTCGCTGCTCCTGCATCCACGCTGGCCGGAGAGCCTTATGTACGTCTATGAGGACAGCGCGGCGGAGAGCGGCATCGGCGGCGGCGGCGGAG  < 100
3' TACTGCTGAAGCGACGAGGACGTAGGTGCGACCGGCCTCTCGGAATACATGCAGATACTCCTGTCGCGCCGCCTCTCGCCGTAGCCGCCGCCGCCGCCTC
          10        20        30        40        50        60        70        80        90

  GAGGAGGCGGCGGCACGGGCGGAGCGGGGGGTGGCTGCAGCGGAGCGAGCCCCGGCAAAGCCCCGAGCATGGATGGTCTGGGCAGCAGCTGCCCGGCCAG  < 200
  CTCCTCCGCCGCCGTGCCCGCCTCGCCCCCCACCGACGTCGCCTCGCTCGGGGCCGTTTCGGGGCTCGTACCTACCAGACCCGTCGTCGACGGGCCGGTC
          110       120       130       140       150       160       170       180       190

  CCACTGCCGCGACCTGCTTCCGCACCCCGTGCTGGGCCGCCCGCCGGCTCCCCTGGGCGCCCCTCAGGGCGCCGTCTATACGGAGATCCCGGCCCCGGAG 3' < 300
  GGTGACGGCGCTGGACGAAGGCGTGGGGCACGACCCGGCGGGCGGCCGAGGGGACCCGCGGGGAGTCCCGCGGCAGATATGCCTCTAGGGCCGGGGCCTC 5'
          210       220       230       240       250       260       270       280       290
```

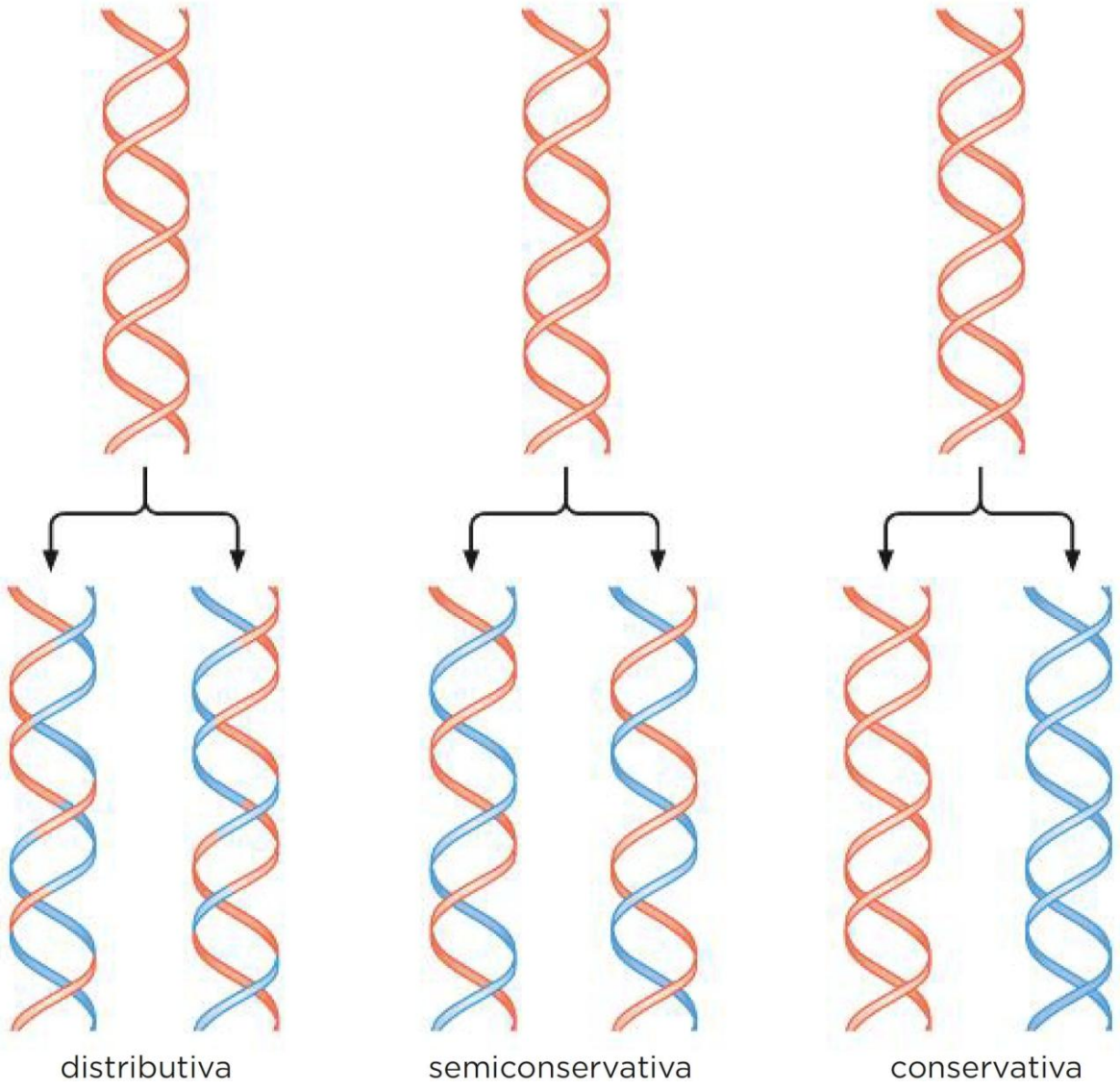
# DNA is replicated

**Why?**

- To **preserve genetic information** across cell generations.
- To allow **cell division and growth**.
- To ensure **accurate inheritance** of traits.

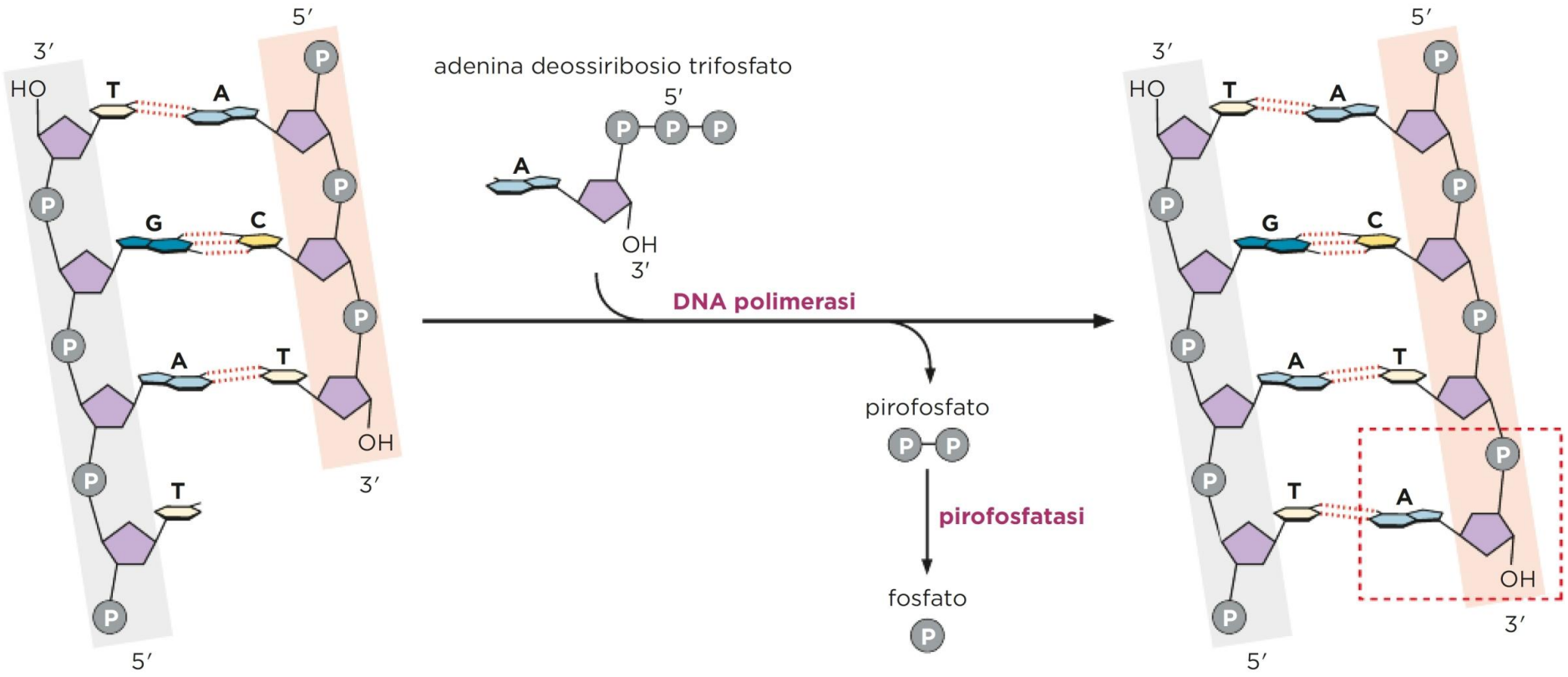
**Different possibilities for DNA replication.**

At the beginning scientists did not know the mechanism by which DNA replication occur.
Experiments clarified that this occurs by a **semi-conservative mechanism**: each strand constitutes a template for the synthesis of complementary strand.






ZANICHELLI

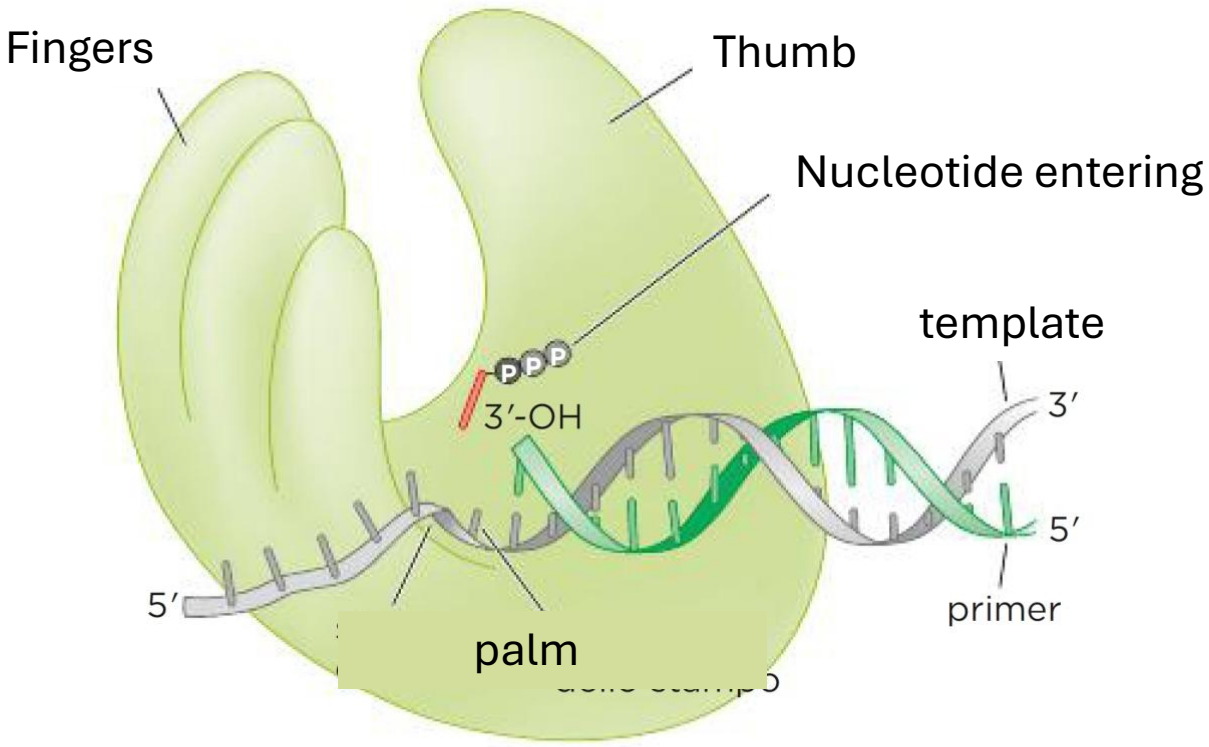



**Enzymatic synthesis of a DNA strand, catalysed by DNA polymerase.** The figure shows the reaction of adding a nucleotide to a growing DNA strand catalysed by DNA polymerase. Although DNA polymerase can catalyse DNA synthesis on its own, in the cell the pyrophosphate molecule released is converted into two phosphates by an enzyme called pyrophosphatase, further driving the nucleotide addition reaction.



ZANICHELLI

# DNA polymerase




The DNA polymerase is a protein (enzyme) with the shape of a partially closed right hand that:

- Interacts with the DNA and slides onto DNA while syhtesizing new strand
- Keeps the primer and the active site in the correct position for DNA synthesis
- Stabilizes the substrate–dNTP complex that drives polymerization

# DNA polymerase substrates

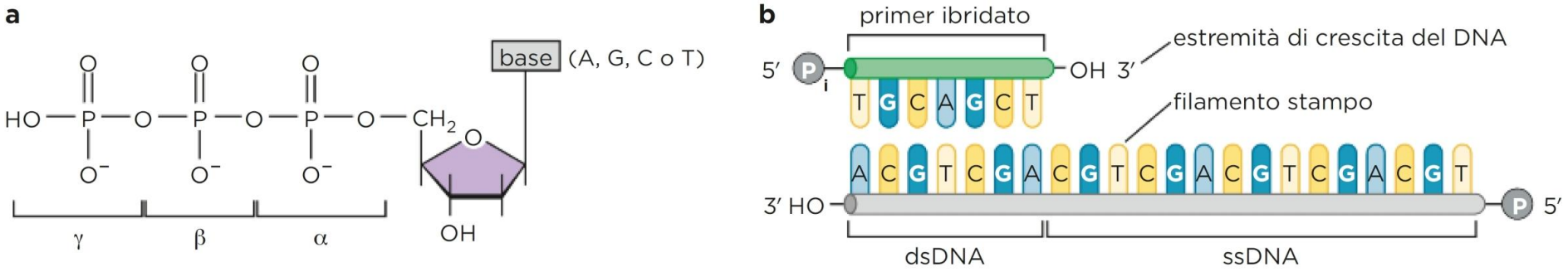



**FIGURA 9.1 I substrati necessari per la sintesi del DNA.** (**a**) La struttura generale del 2′-deossinucleoside trifosfato. Sono indicate le posizioni dei fosfati α, β, e γ. (**b**) Struttura di una giunzione primer:stampo. La corta molecola del primer è completamente ibridata con il filamento di DNA più lungo e deve avere un'estremità 3′-OH adiacente alla regione a singolo filamento dello stampo. Il filamento di DNA più lungo comprende una parte ibridata con il primer e una porzione di singolo filamento che funziona come stampo per la nuova sintesi. Il DNA neosintetizzato si estende dal terminale 3′ del primer.

**2'-Deoxy-nucleotides triphosphate**

dGTP
dATP
dTTP
dCTP

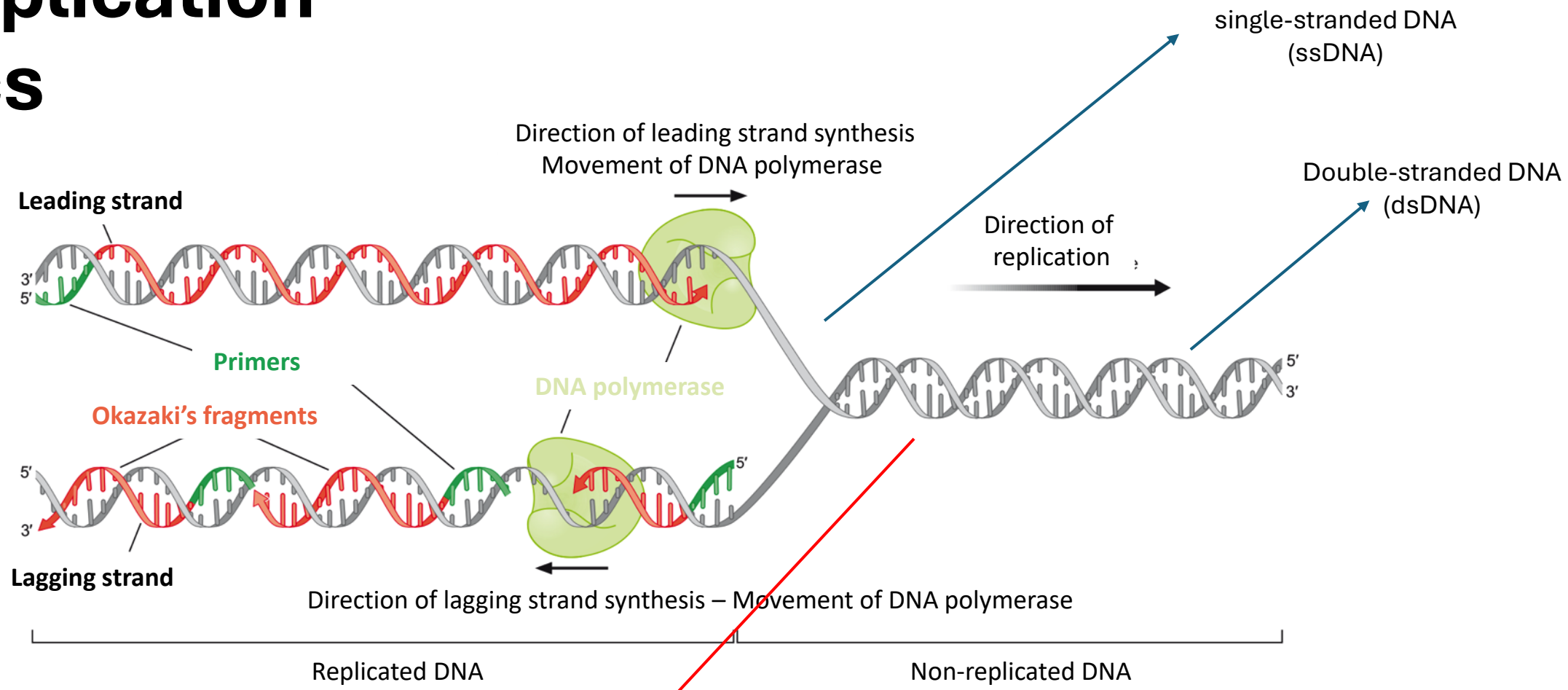
**Primer hybridized to template**

DNA polymerase never catalyses the synthesis of a polynucleotide de novo

DNA synthesis always occurs in the 5′→3′ direction

The primer has a complementary sequence targeting upstream the sequence to be amplified

# DNA replication logistics




DNA unwinding requires energy!

- In the cells: proteins acting as motors
- In the lab: high temperature! (>90°C), this process is also known **denaturation**

In both cases it is a reversible process because hydrogen bonds are not covalent

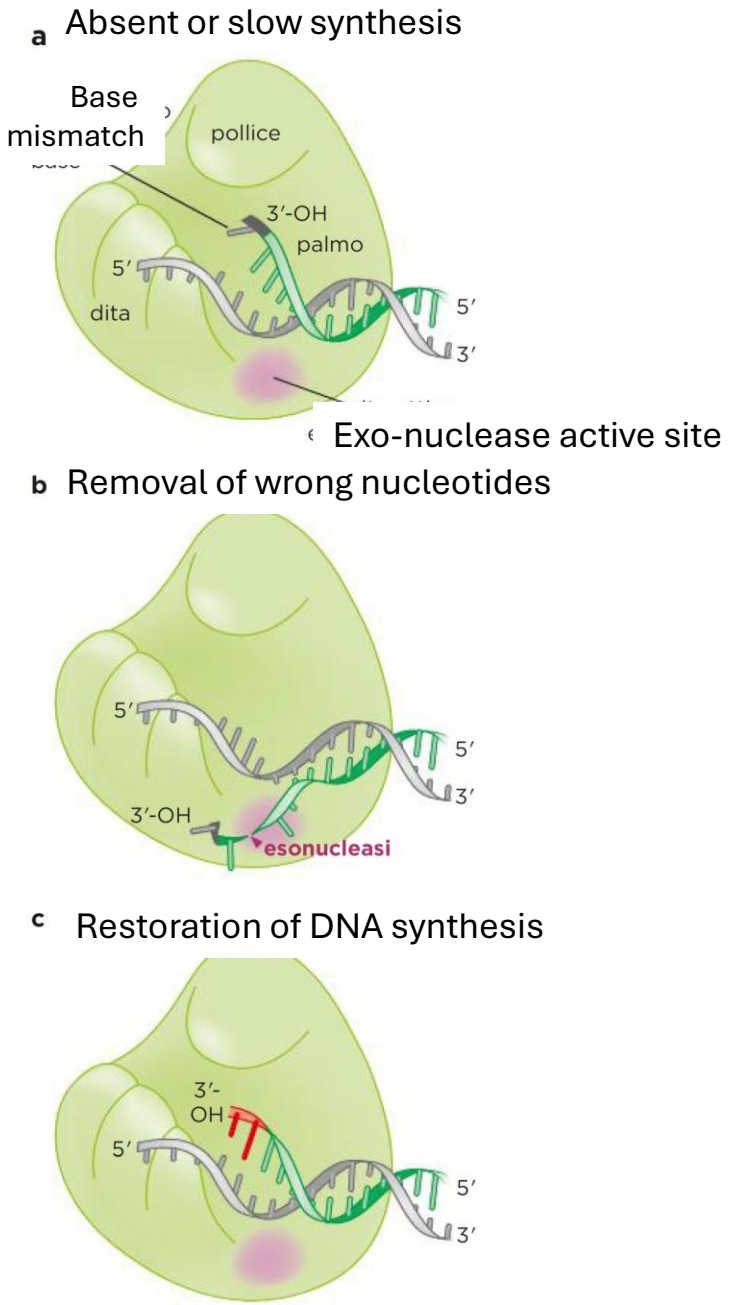
# DNA polymerase
# Proof-reading activity

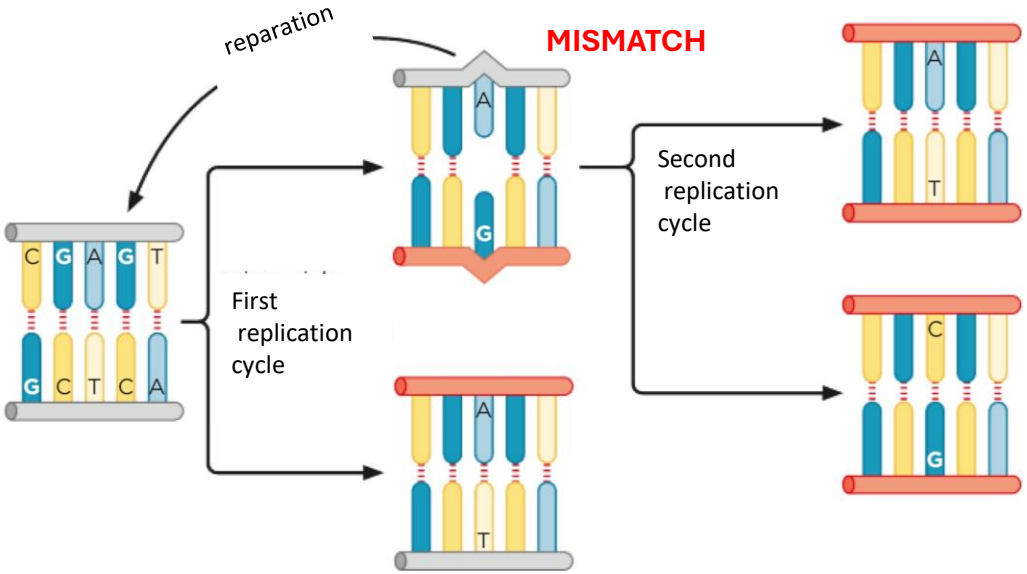
(a) When an incorrect nucleotide is incorporated into DNA, the rate of synthesis decreases due to the improper positioning of the 3'-OH.

(b) When there is a mismatched 3' end, the last 3–4 nucleotides of the primer become single-stranded, which increases their affinity for the exonuclease active site. Once bound, the noncomplementary nucleotide (and often one additional nucleotide) is removed from the primer.

(c) After the incorrect nucleotide has been excised, a correctly base-paired primer–template junction is re-formed, and DNA synthesis resumes (the newly synthesized DNA is shown in red).

- DNA polymerase inserts **1 wrong dNTP every $10^5$ nucleotides**
- Proofreading exonuclease reduces errors to **1 in $10^7$**
- Final mutation rate per genome is **~1 in $10^{10}$**, thanks to **additional repair mechanisms**






Example of mutation introduced during replication.

Mutations can be introduced also on already synthesized DNA (e.g. UV radiation or chemical damage)

**What is a mutation?**

- A **permanent change** in the DNA sequence.
- Can affect a single nucleotide or large genomic regions.

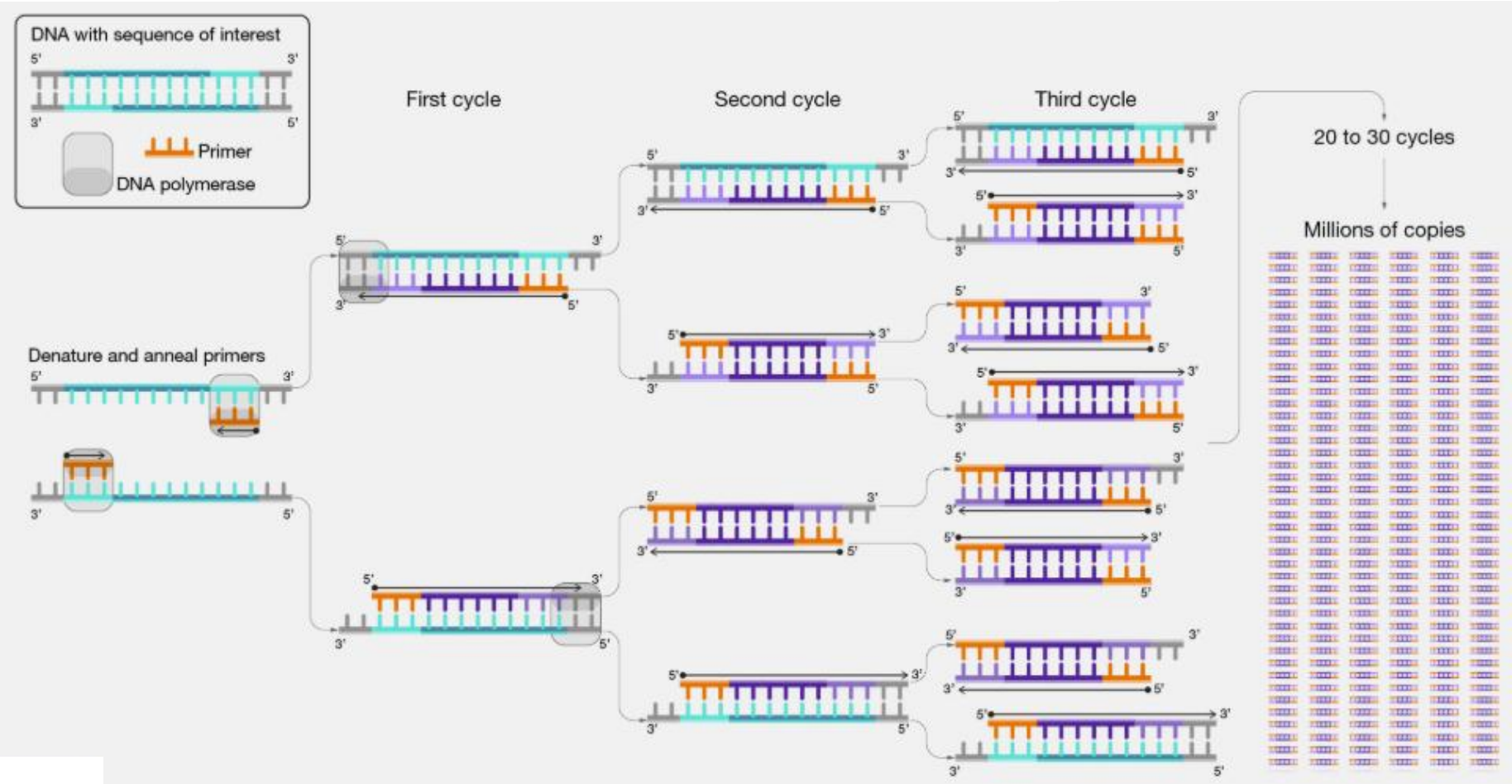
**Mutations in the human genome**

- Each human genome carries **~60–100 new mutations** per generation.
- Types: **point mutations, insertions, deletions, chromosomal rearrangements**.

**Mutations vs. Polymorphisms**

- **Mutation**: rare (<1% frequency), may cause disease.
- **Polymorphism**: common (>1% frequency), usually neutral or benign (e.g., **SNPs**= single nucleotide polymorphisms).

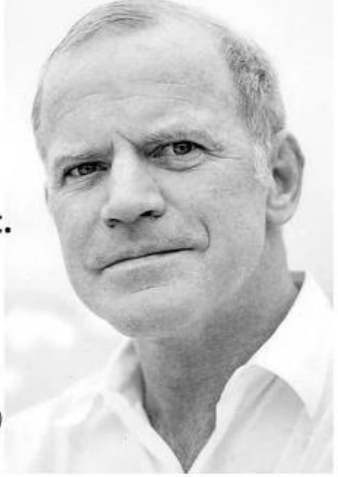
# Polymerase chain reaction (PCR)




"What if I had not taken LSD ever; would I have still invented PCR?

I don't know. I doubt it. I seriously doubt it."

*- Kary Mullis*
*(Winner of the 1993 Nobel Prize in Chemistry for his discovery of the polymerase chain reaction, or PCR)*



- DNA template
- DNA polymerase

- Primers (FW + RV)
- DNA polymerase

- Thermocycler performing the 94°-60°-72° cycle

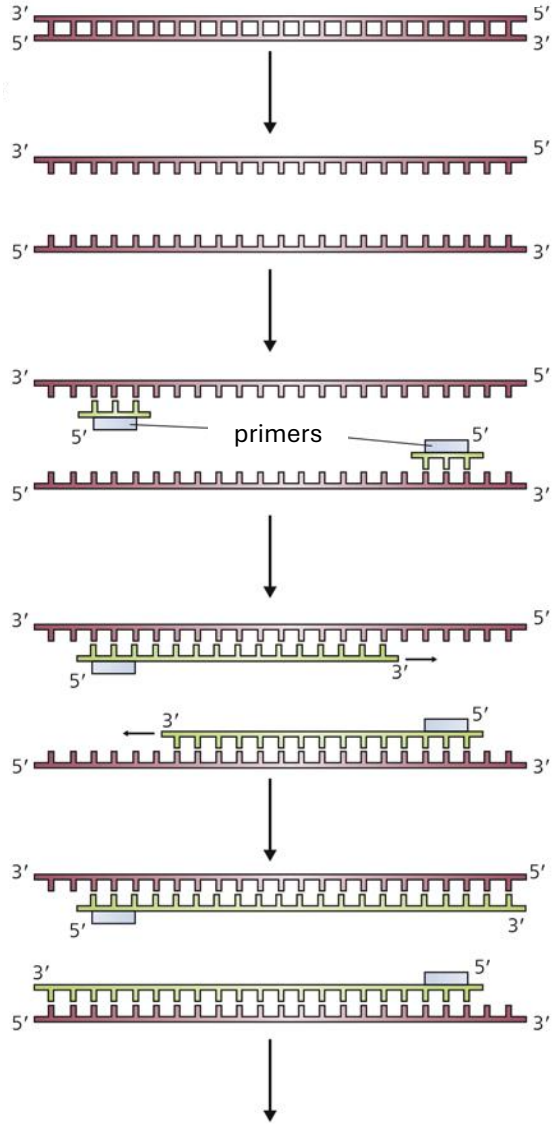
# Scheme of a PCR reaction

Double strand DNA template

1. Denaturation of template at 94°C

2. Primer annealing at 50°-60°C

3. New DNA synthesis at 62-74°C

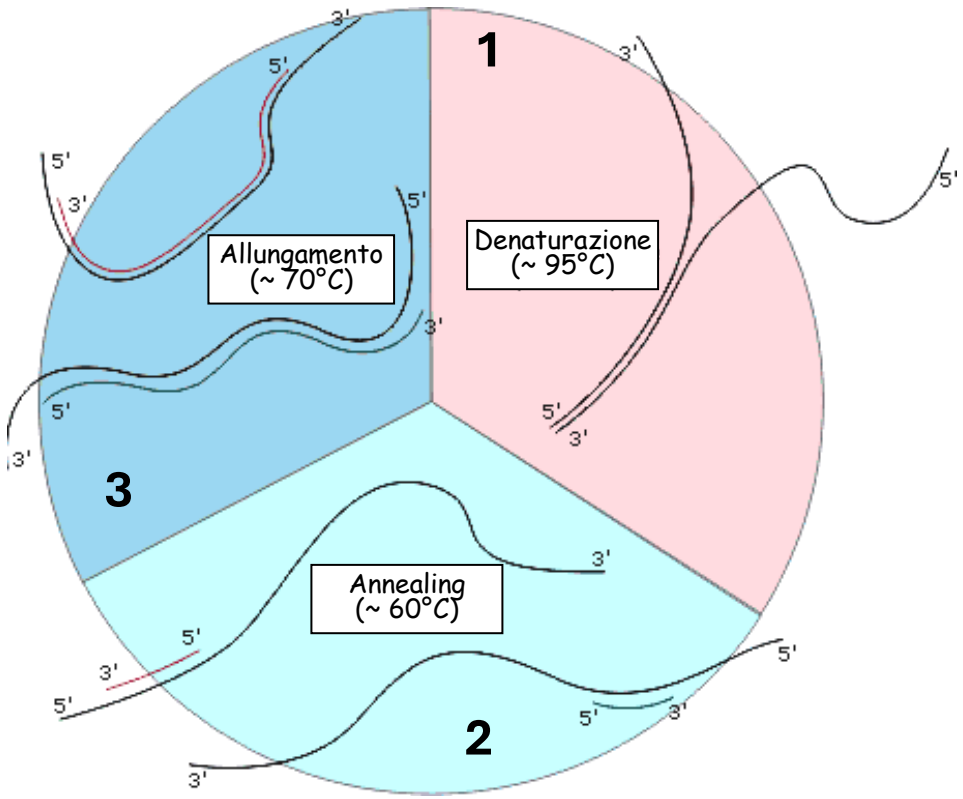



4. Repetition for 25-30 times

## Phases of a PCR cycle

### 1) Denaturation:

DNA is denatured by heating at ~ 92-95°C: the two strands become separated




### 2) Annealing/hybridization:

The mixture is cooled until it reaches the temperature that ensures the specific hybridization of the primers to the complementary regions of the template.

### 3) Extension:

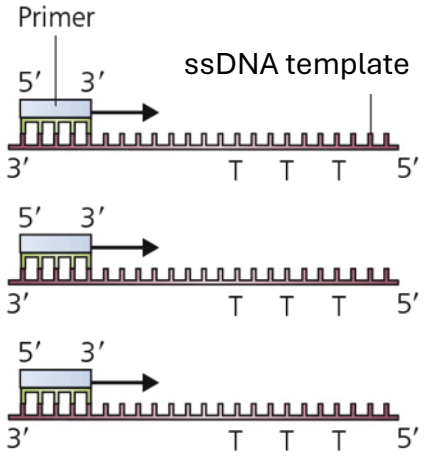
The temperature of the mixture is raised to 68–74°C, allowing the heat-stable DNA polymerase to synthesize the complementary strand of the template starting from the oligonucleotide primer.

# PCR can be exploited for DNA sequencing

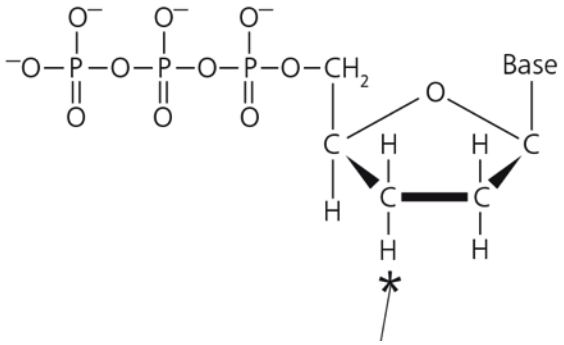
## Sanger sequencing
**(first generation of sequencing technologies)**

- Determine DNA sequence.
- Use template + primer.
- DNA polymerase synthesizes new strand.
- ddNTPs terminate synthesis at specific bases.
- Fragment sizes are used to infer the sequence.

(a) Synthesis start

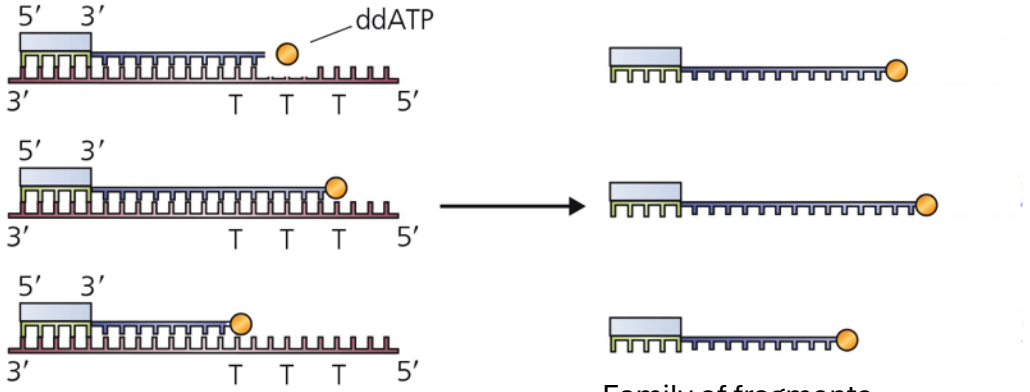



(b) Use of Di-deoxy nucleotides (ddNTPs)



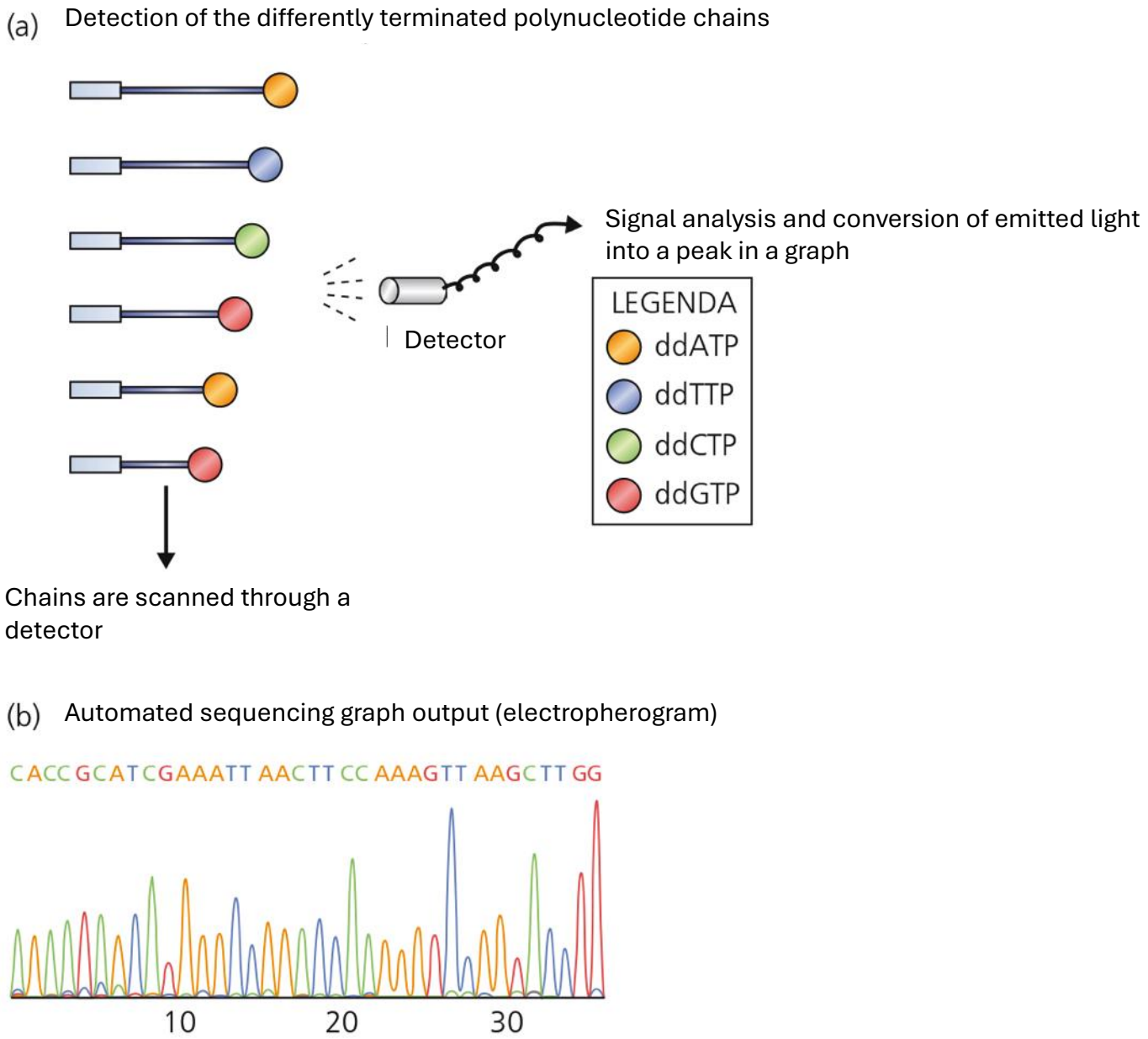

3’ position: -OH substituted with –H, the chain can’t be elongated

(c) Synthesis stops each time a ddNTP is incorporated, at each position

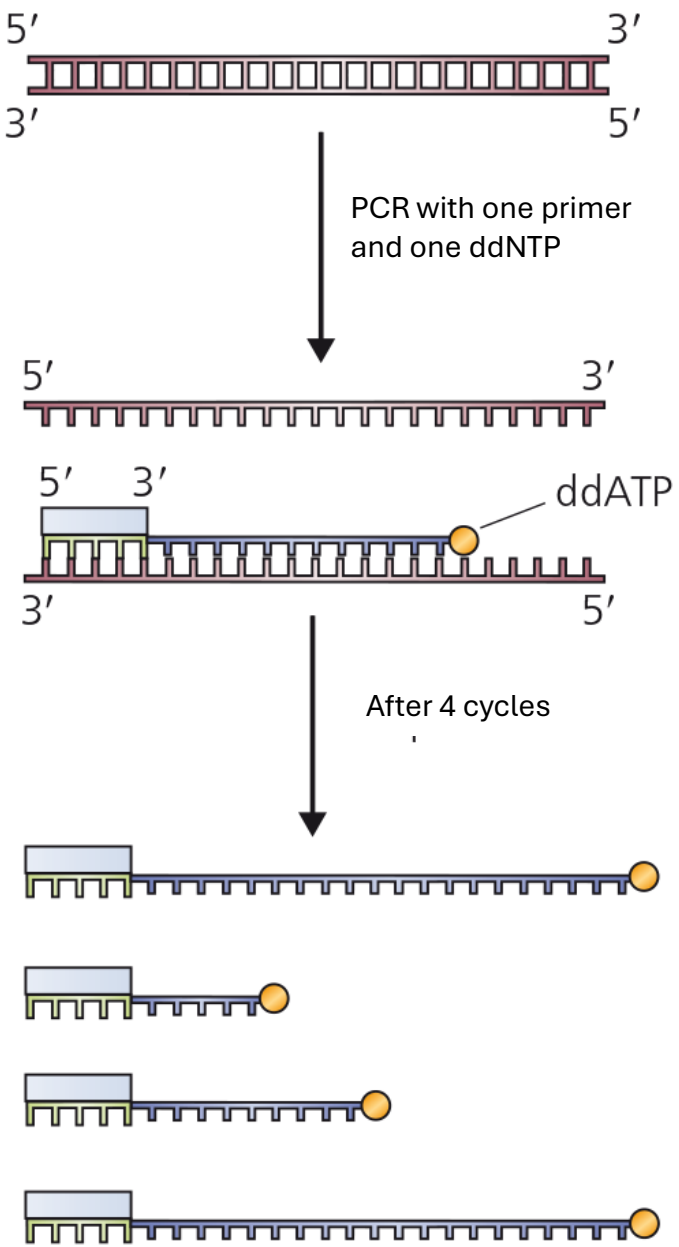



Family of fragments terminated in A

# Throughput increased by color-tagging each ddNTP

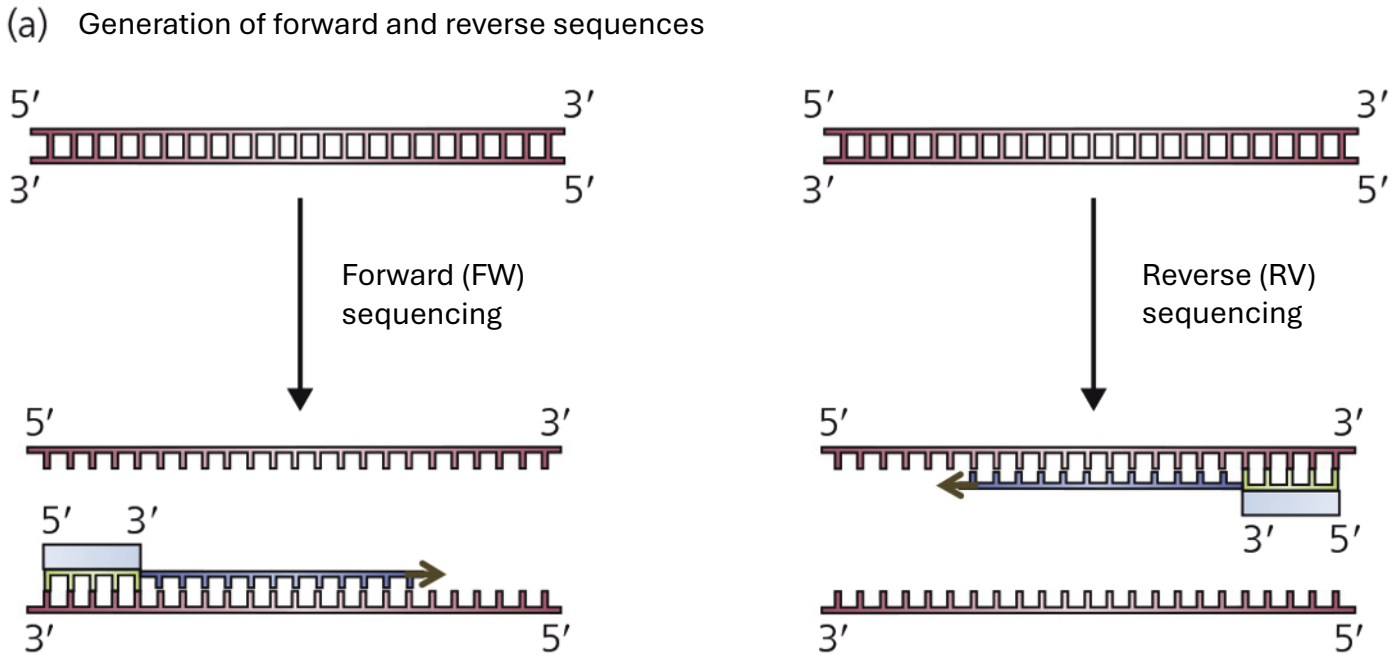


# Sense strand sequencing




Family of polynucleotides interrupted in A.
Assuming to do it for all ddNTPs, one can obtain all interrupted sequences by a number of repeated amplification cycles ensuring to cover the all sequence.

# Sense and antisense strand sequencing




As the two strands contain the same information, FW+RV sequencing is a way to increase the number of sequences obtained, besides amplification

→ sequencing precision

# Read alignment

```
                                AGCATCGTAGCTTCAGTATGATGATGCTAG   Read   1
ATGATCGTAGCTAGCATCGTAGCTAGC                                      Read   2
   ATCGTAGCTAGCATCGTAGCTAGCATCGTAGCTT                            Read   3
                                TTGTAGCTTCAGTATGATGATGCTAG       Read   4
              GCATCGTAGCTAGCATCGTAGCTTCAGT                       Read   5
ATGATCGTAGCTAGCATCGTA                                            Read   6
ATGATCGTAGCTAGCATCGTAGCTAGCATCGTAGCTTCAGTATGATGATGCTAG   Deduced sequence
```

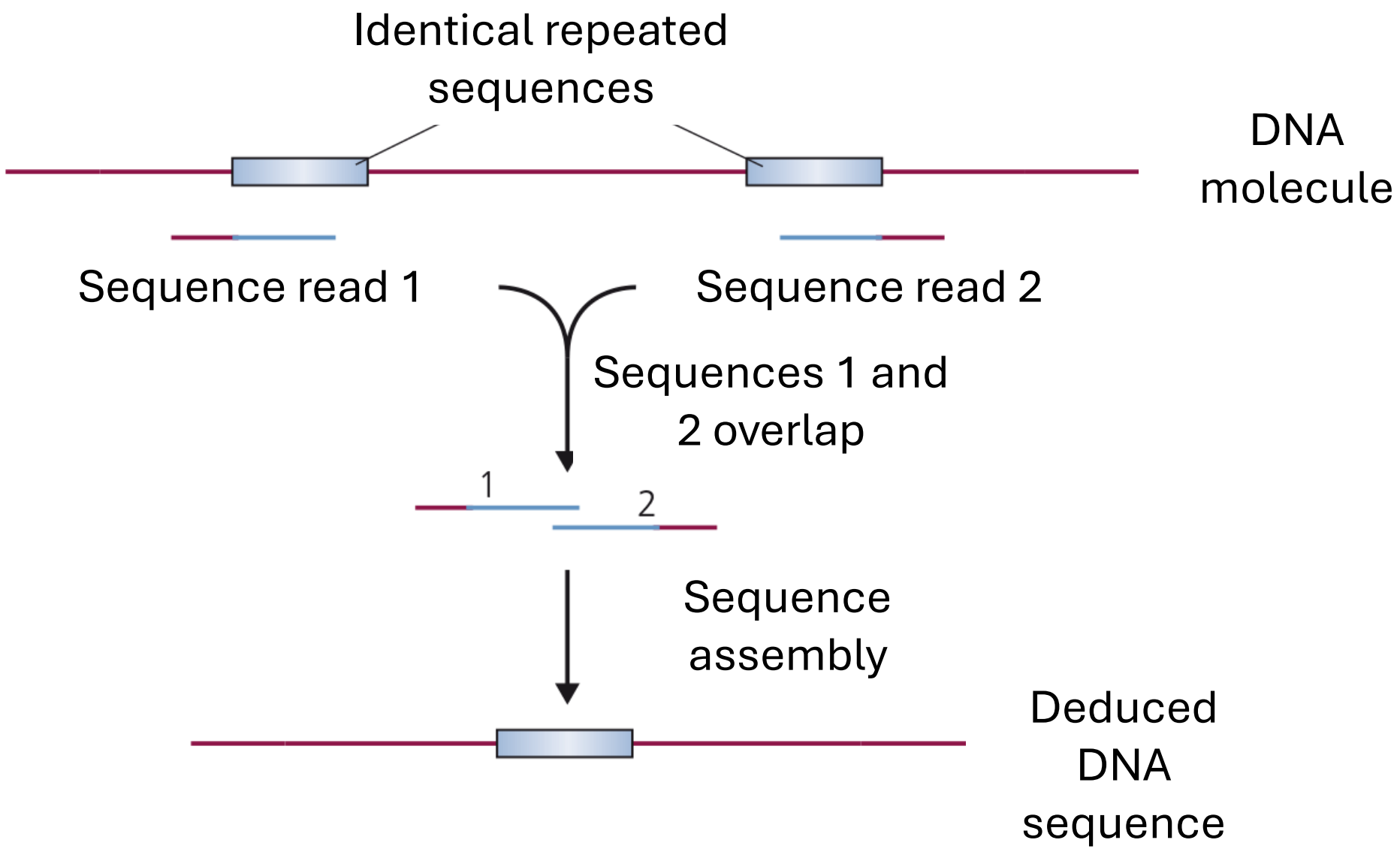
For each sequence of interest, the sequence is derived several times in order to identify errors present in the single reads. In this example, the grey column underlines a discrepancy in read 4 that can be attributed to a sequencing error.

→ Sequencing precision

Furthermore, the length of a read is often shorter than the length of the desired gene sequence. Therefore, several partially overlapping sequences are used to derive the entire sequence of interest.

→ Sequence assembly

# Challenges in read alignment



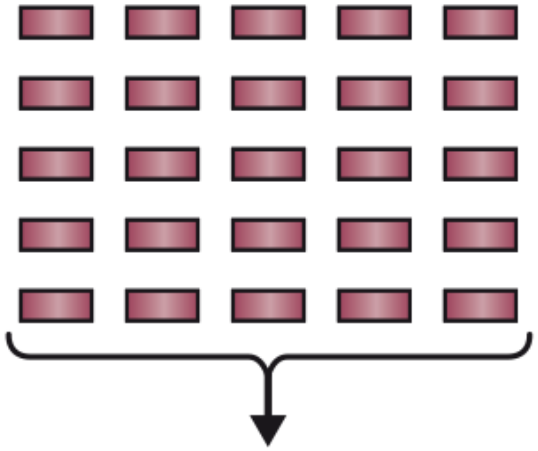
# First generation sequencing

# Second generation sequencing

**Next generation sequencing**

Single DNA sequence, typically isolated by PCR ($\approx 10^3$ nt long)

TACGGATGGCATAGA

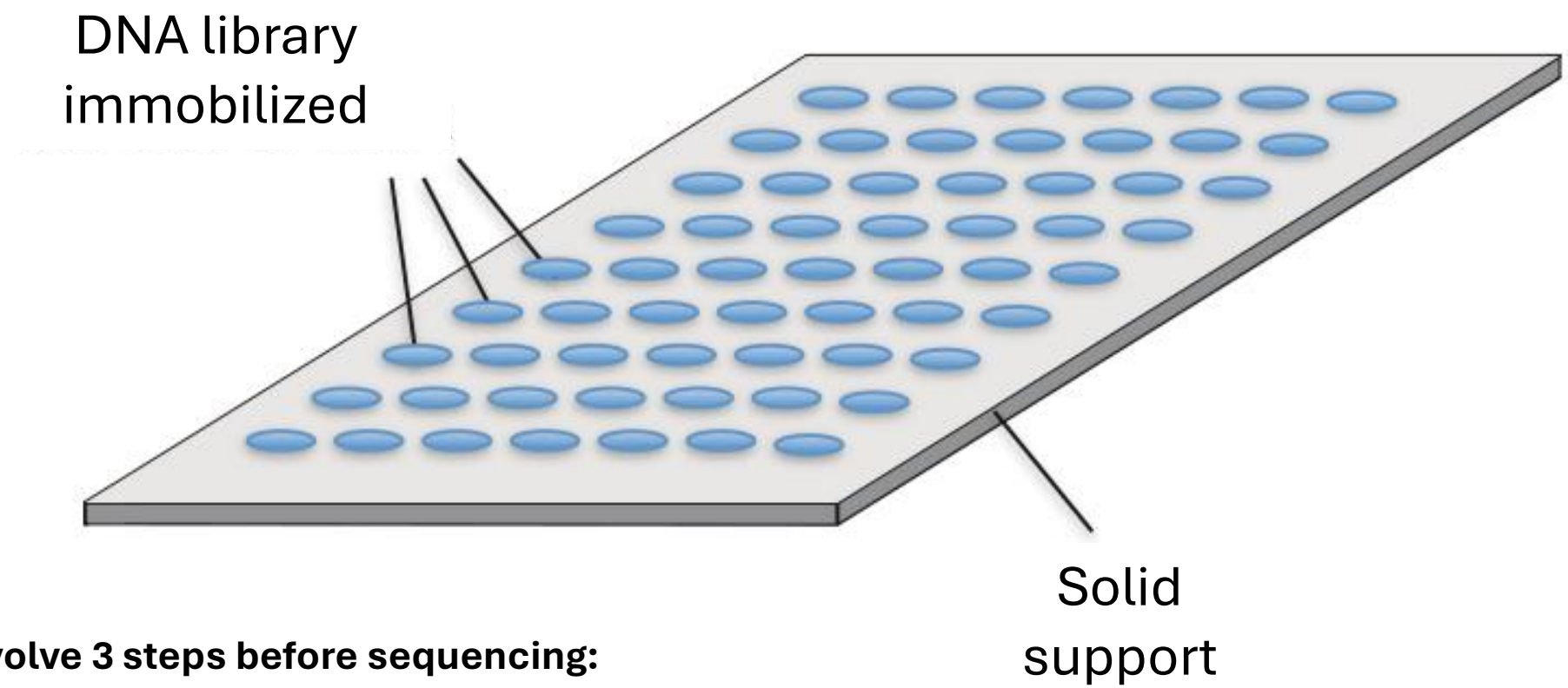


Array of different DNA sequences, typically obtained by fragmentation ($\approx 10^2$ nt long)

ACGTATCATGCGGATGG
TAGCATGACGTAGCGTT
GTAGCAGGTACGATGCC
GTAGACGATGCAGCATC
TAGGACCTAGCCGGACA

**Output**: one sequence, up to $10^3$ nt long

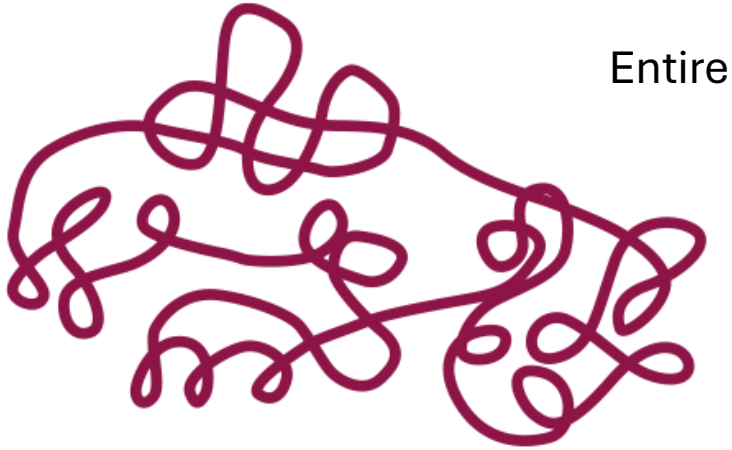
**Output**: many shorter sequences ($10^6$-$10^9$) up to $\approx 10^2$ nt long




**All NGS methods involve 3 steps before sequencing:**

1. **Fragmentation** of DNA into sizes suitable for the sequencing method
2. **Immobilization** of the fragments on a solid support
3. **Amplification** of the immobilized fragments

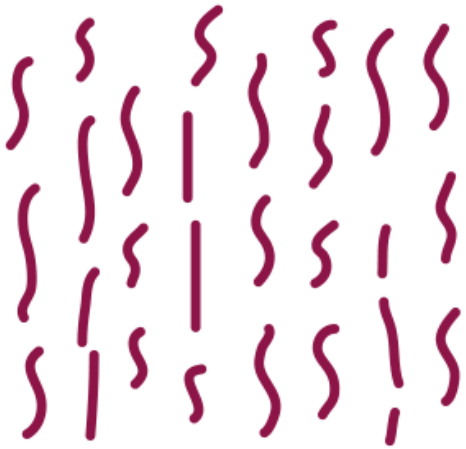
In this way, the sequencing reactions are carried out on the pool of amplified samples in parallel in a matrix format.



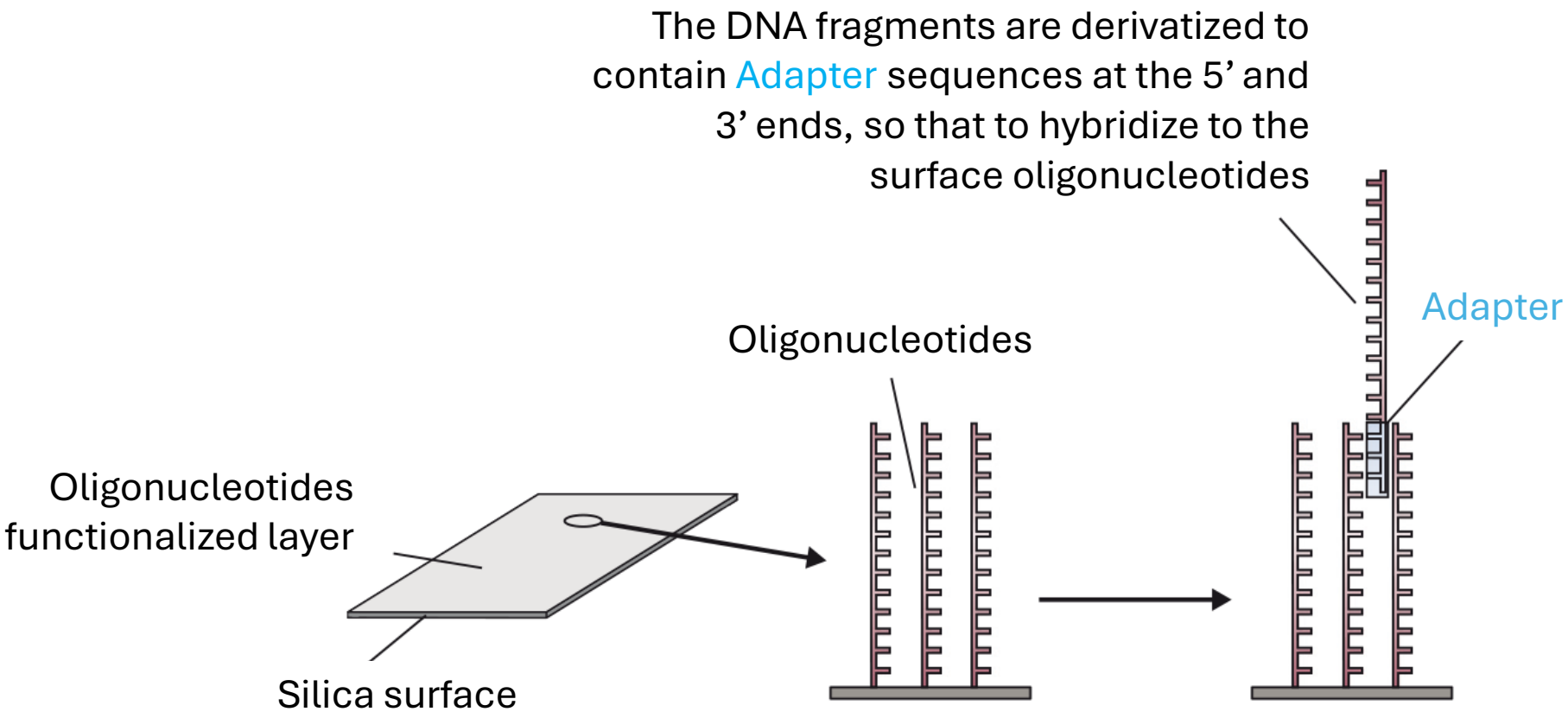
Entire genome corresponding to 46 very long dsDNA molecules

Sonication

Sonication generates random fragments ranging from **50 to 500 nt**. In this way, portions of the entire genome are immobilized on the array.



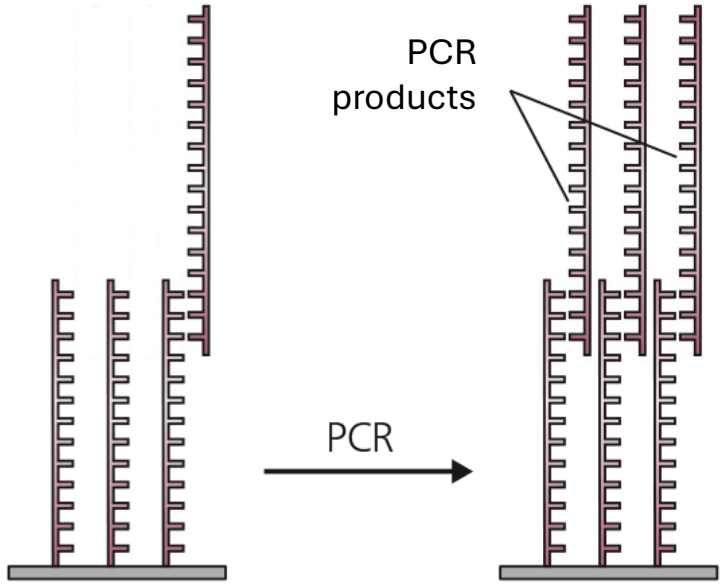
Short random fragments




**Immobilization** of DNA library fragments through hybridization to oligonucleotides fixed on a glass support.

**NB:** the Adapter sequence is used to hybridize the primer, thus amplified and sequenced. Typically, it is a sequence conserved throughout all sequences that is eliminated before proceeding to analysis.

**Library amplification**

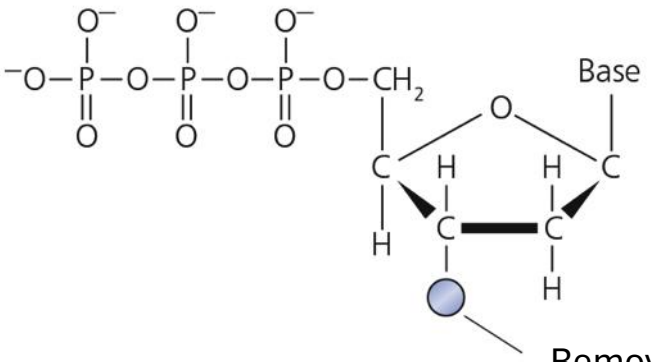



The newly amplified PCR products immobilize to the oligonucleotides adjacent to the one bound to the template, thereby generating a group of identical fragments called a CLUSTER.

→ Cluster generation allows the physical separation of different DNA amplicons, allowing the array to be formed

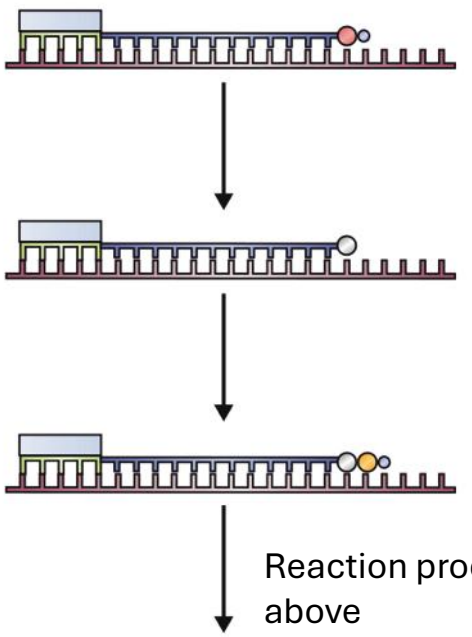
# Illumina sequencing

(a) Reversible terminator nucleotide




Sequencing of max 300 bp for each fragment....but the possibility of doing it in parallel for many fragments guarantees the reading of 2000 Mbp for each sequencing

(b) Sequencing by reversible terminator nucleotide incorporation



Based on colour-tag, the terminal nucleotide is identified as G

The blocking group and the colour-tag are both removed

The next terminal nucleotide is incorporated and identified as A

Reaction proceeds as above

Illumina Sequencing Technolgoy

https://www.youtube.com/watch?v=fCd6B5HRaZ8