Doctor 022

Biochemistry

Sheet no. 5

DNA SEQUENCING

Writer: Al-Razi Node Team

Corrector: Al-Razi Node Team

Doctor: Dr.Mamoun, Dr.Diala

WHAT IS DNA SEQUENCING?

DNA sequencing is the process of determining the exact order of nucleotides in a genome or in a DNA fragment.

❖ The importance of knowing the DNA sequence:

- **Identification of genes and their localization:** so you can determine where gene x is located on the arm of chromosome, if it's near to the centromere or telomere, also identify were these telomeres and centromeres are, and exons by localizing the sequence they begin and end with.
- **Identification of protein structure and function:** when we identify the gene, we can identify the codons on the gene's RNA and translate them to amino acids, so we would be able to know the sequence of the proteins then predict what the structures, functions and localization of these proteins, also what proteins it can interact with.
- **Identification of DNA mutations:** we can compare the sequence of our unknown DNA (or DNA from someone with a certain disease) to the database for normal human genome, so we can pinpoint where exactly a mutation occurs, and how it's related to a certain disease.
- **Genetic variations among individuals in health and disease:** by knowing the sequence of the DNA of an individual, we would be able to know how variable it is, in comparison with other individuals.
- **Prediction of disease-susceptibility and treatment efficiency:** we will be able to predict if a person may experience some kind of illness in the future that is carried on by a family gene, which we also see in other family members (مثلا بنلاقي افراد من العيله صار معهم نفس المرض بنفس الوقت المتوقع تقريبا).
- **Evolutionary conservation among organisms:** determine how organisms are related to each other. We use animals as model systems, we compare human genes with some animal genes (mouse, cats,

dogs…etc) in order to help us understand how our cells function normally, or whenever there is a mutation.

DNA SEQUENCING OF ORGANISM GENOME

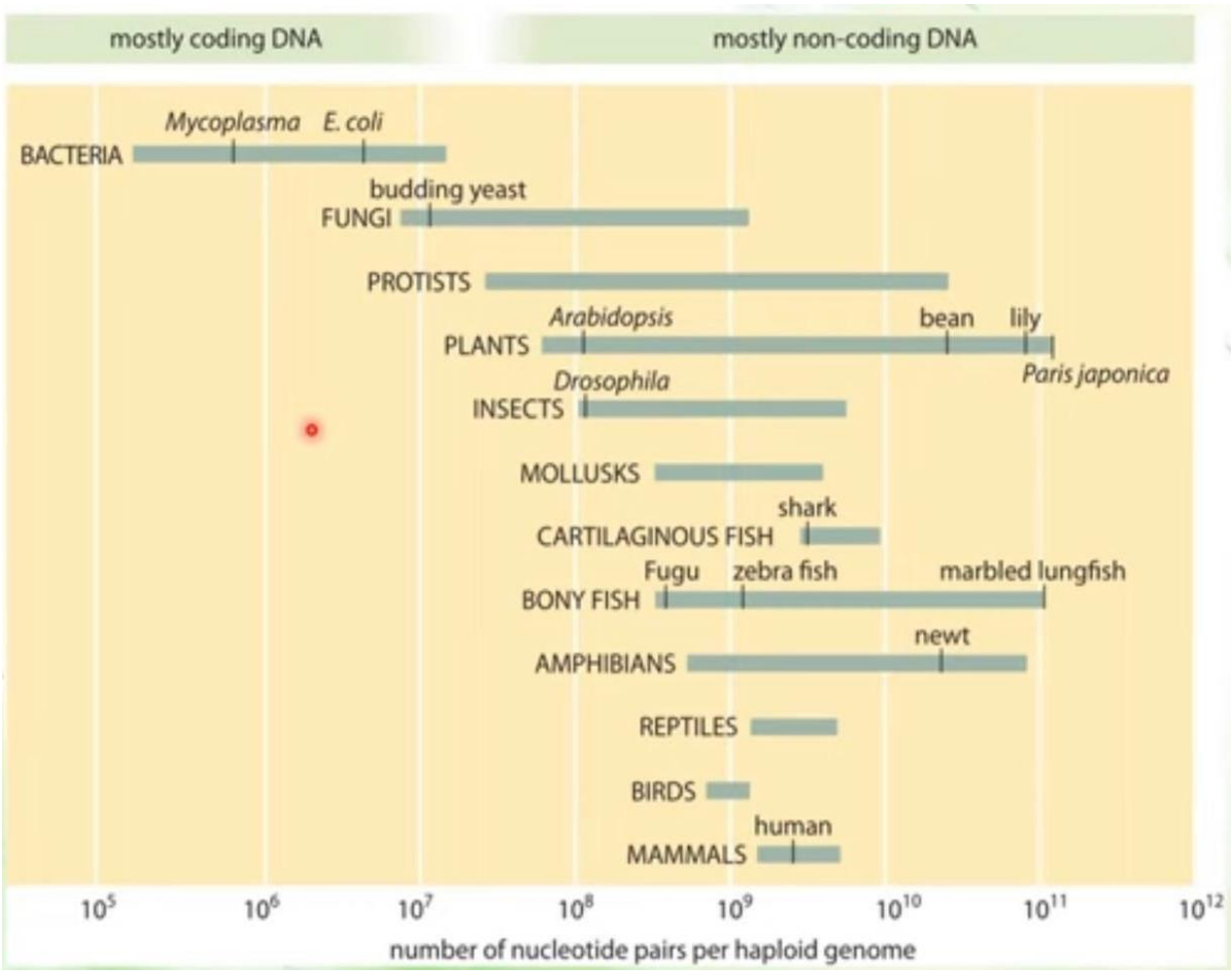
- **Viruses and prokaryotes first**
- **Human mitochondrial DNA**
- **The first eukaryotic genome sequenced was that of yeast, Saccharomyces cerevisiae.**
- **The genome of a multicellular organism, the nematode Caenorhabditis elegans.**
- **Determination of the base sequence in the human genome was initiated in 1990.** (it hasn't been completed yet, we still have to determine the final regions of **Y** chromosome)

---

organism	genome size (base pairs)	protein coding genes	number of chromosomes
model organisms			
model bacteria *E. coli*	4.6 Mbp	4,300	1
budding yeast *S. cerevisiae*	12 Mbp	6,600	16
amoeba *D. discoideum*	34 Mbp	13,000	6
nematode *C. elegans*	100 Mbp	20,000	12 (2n)
fruit fly *D. melanogaster*	140 Mbp	14,000	8 (2n)
model plant *A. thaliana*	140 Mbp	27,000	10 (2n)
mouse *M. musculus*	2.8 Gbp	20,000	40 (2n)
human *H. sapiens*	3.2 Gbp	21,000	46 (2n)
viruses			
hepatitis D virus (smallest known animal RNA virus)	1.7 Kb	1	ssRNA
HIV-1	9.7 kbp	9	2 ssRNA (2n)
influenza A	14 kbp	11	8 ssRNA
bacteriophage λ	49 kbp	66	1 dsDNA
organelles			
mitochondria - *H. sapiens*	16.8 kbp	13 (+22 tRNA +2 rRNA)	1
chloroplast – *A. thaliana*	150 kbp	100	1
eukaryotes - multicellular			
dog *C. familiaris*	2.4 Gbp	19,000	40
chimpanzee *P. troglodytes*	3.3 Gbp	19,000	48 (2n)

- This table shows a comparison of the genomes of different organisms, between protein coding genes with the number of chromosomes. (look at these numbers and enjoy, do not **memorizzze** them ☺)

NUCLEOTIDES PER GENOMES

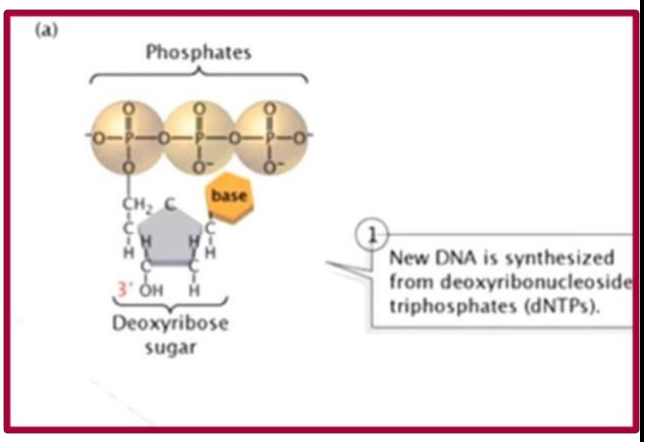


- This graph shows comparison of genomes of different organisms, you can see variable genome sizes in different species **(again do not memorize).**

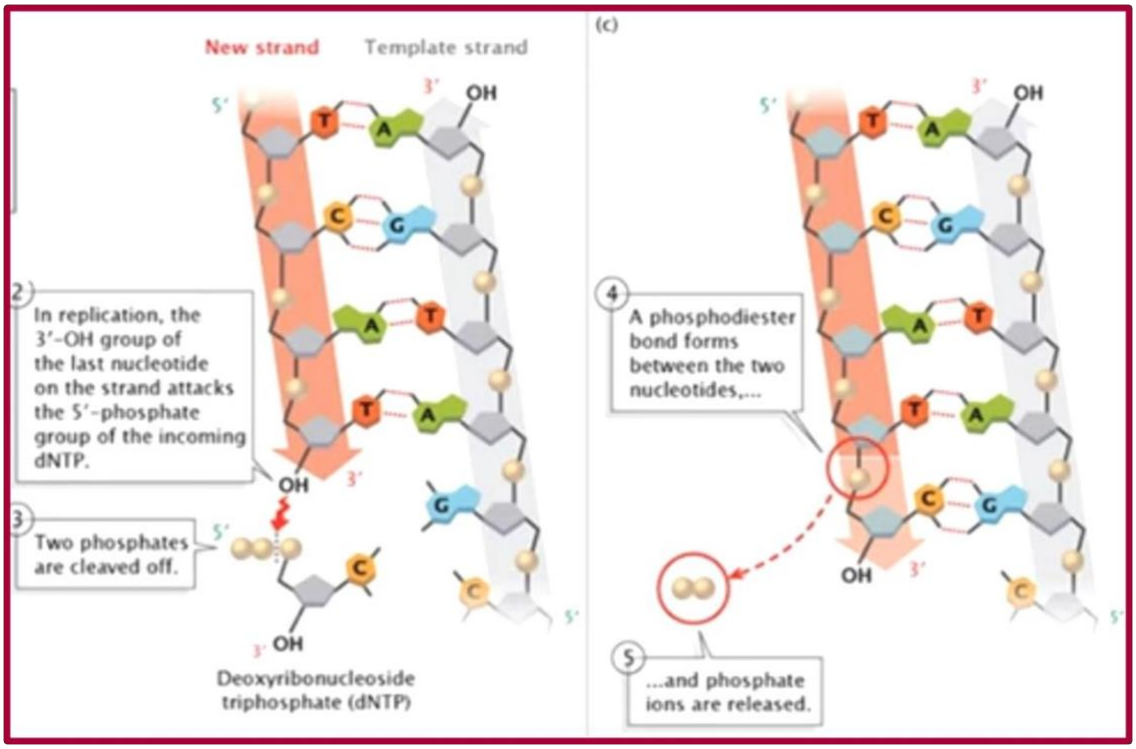
DNA SYNTHESIS/ELONGATION

Reminder:

- ❖ nucleotide structure:
  - ➢ deoxy ribose sugar (because it's missing a hydroxyl group at carbon no.2)
  - ➢ triphosphate group (at carbon no.5)
  - ➢ nitrogenous base (at carbon no.1)



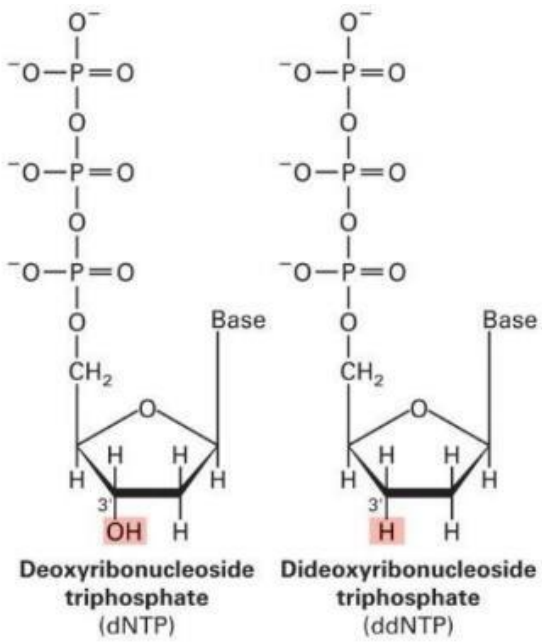
- ❖ In order to synthesize DNA, we need a triphosphate nucleotide at the 5' and hydroxyl group at the 3' end.
- ❖ So we have phosphodiester bond formation between carbon no.3 with carbon no.5 in the incoming nucleotide. So we get energy from the release of two phosphate group.



METHOD OF DNA SEQUENCING

The most popular method is based on premature termination of DNA synthesis by dideoxynucleotides.

- Dideoxynucleotide basically is similar to the deoxy nucleotide (sugar, base, three phosphate groups) but it differs that it misses a hydroxyl group at carbon no.3.
- When we add this molecule to the DNA, no other nucleotide can be added to the 3' end and there is no formation of phosphodiester bond (because there is no hydroxyl group).
- As a result, DNA synthesis stops at this point (this will be the last nucleotide that is added to the DNA).



THE PROCESS...

1- DNA synthesis is initiated from a primer that has been labeled with a radioisotope.

In order to sequence a DNA, we need a primer, DNA polymerase, a template and substrates. The primer is labeled with radioisotope like radioactive phosphorus. And the substrates would obviously include (deoxyAtriphosphate, dGtp, dTtp, and dCtp), ut here we also add dideoxy(A,T,C,G)triphosphate that is labelled with a florescent color. (So we can tell were it is on the newly synthesized DNA strand)

2- Separate reactions are run, each including deoxynucleotides plus one dideoxynucleotide (either A, C, G, or T)

Lets say we have 1000 strands to be synthesized, keep in mind that the amount of the usual added nucleotides (dntp) highly exceeds the

amount of the dideoxynucleotides, so chances that the usual ones get added to the strand are a lot higher than these of adding the dideoxynucleotides to it, and this whole addition procedure is a random one, (يعني يضاف الdideoxy باي وقت لكلstrand)resulting in different lengths in the new DNA strands.

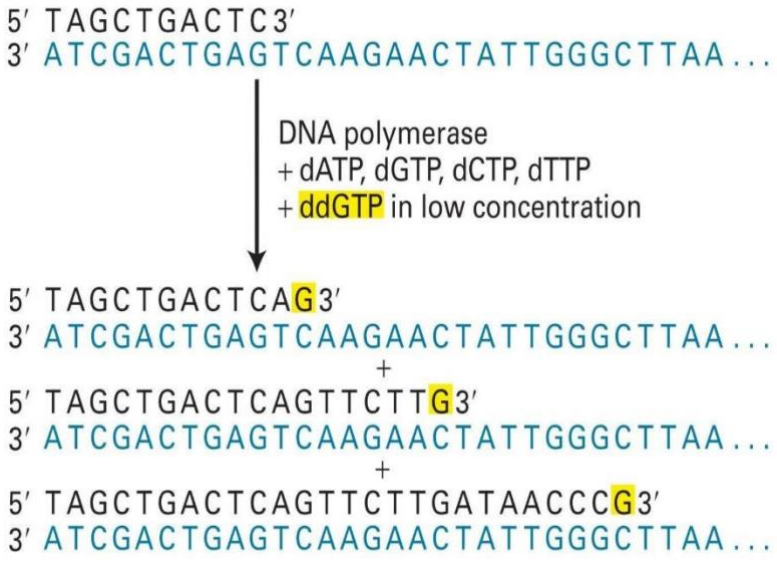
DNA polymerase will start synthesizing DNA, but if it adds a deoxyribonucleotide it can continue, but if it adds a

Dideoxyribonucleotide, synthesis is terminated (because no other nucleotide can be added to the deoxy 3' end of the DNA).

3- Incorporation of a dideoxynucleotide stops further DNA synthesis because no 3 hydroxyl group is available for addition of the next nucleotide.

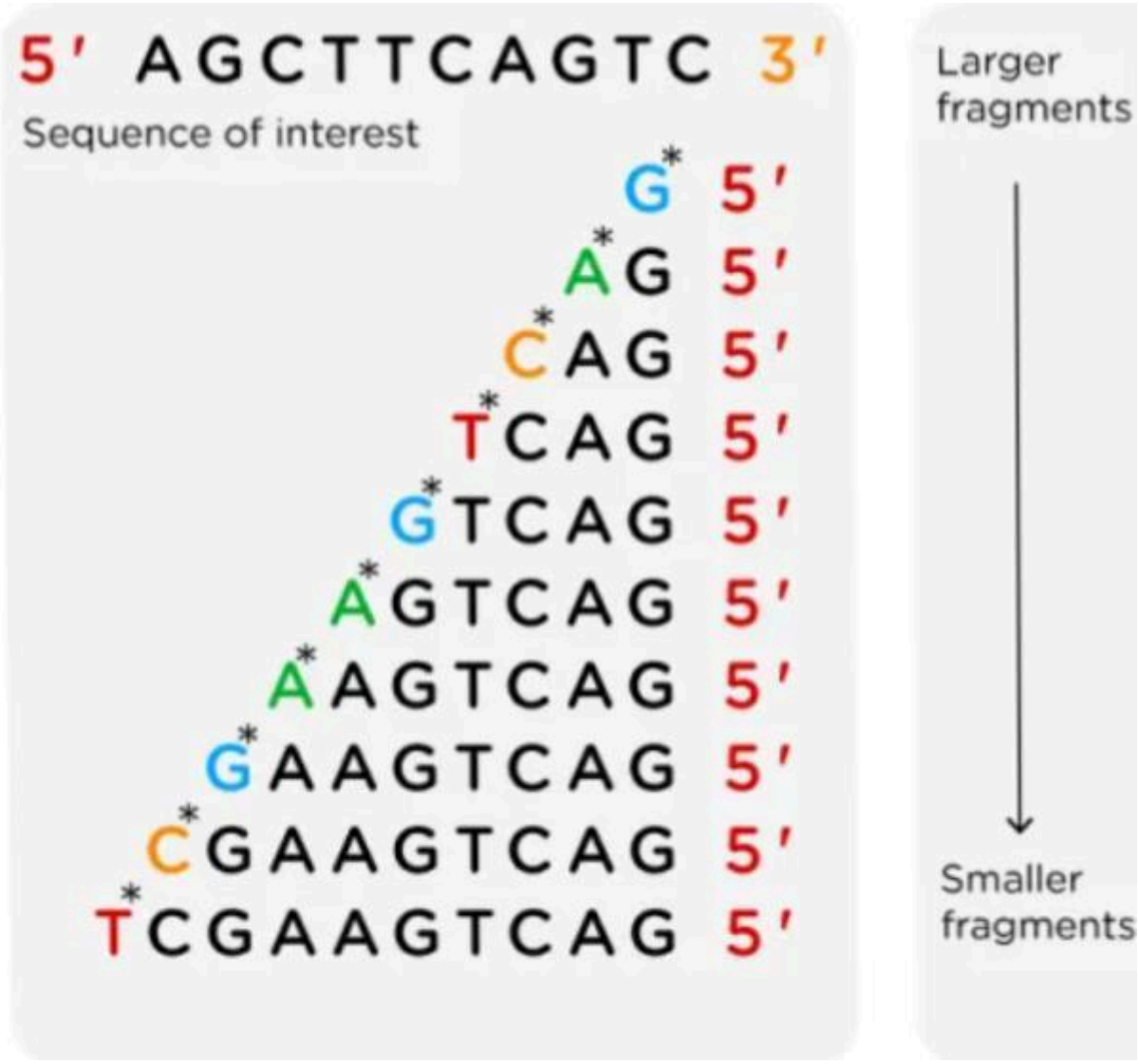
GENERATION OF FRAGMENTS

- **A series of labeled DNA molecules are generated, each terminated by the dideoxynucleotide in each reaction.**
- **These fragments of DNA are then separated according to size by gel electrophoresis and detected by exposure of the gel to X-ray film.**
- **The size of each fragment is determined by its terminal dideoxynucleotide, so the DNA sequence corresponds to the order of fragments read from the gel.**



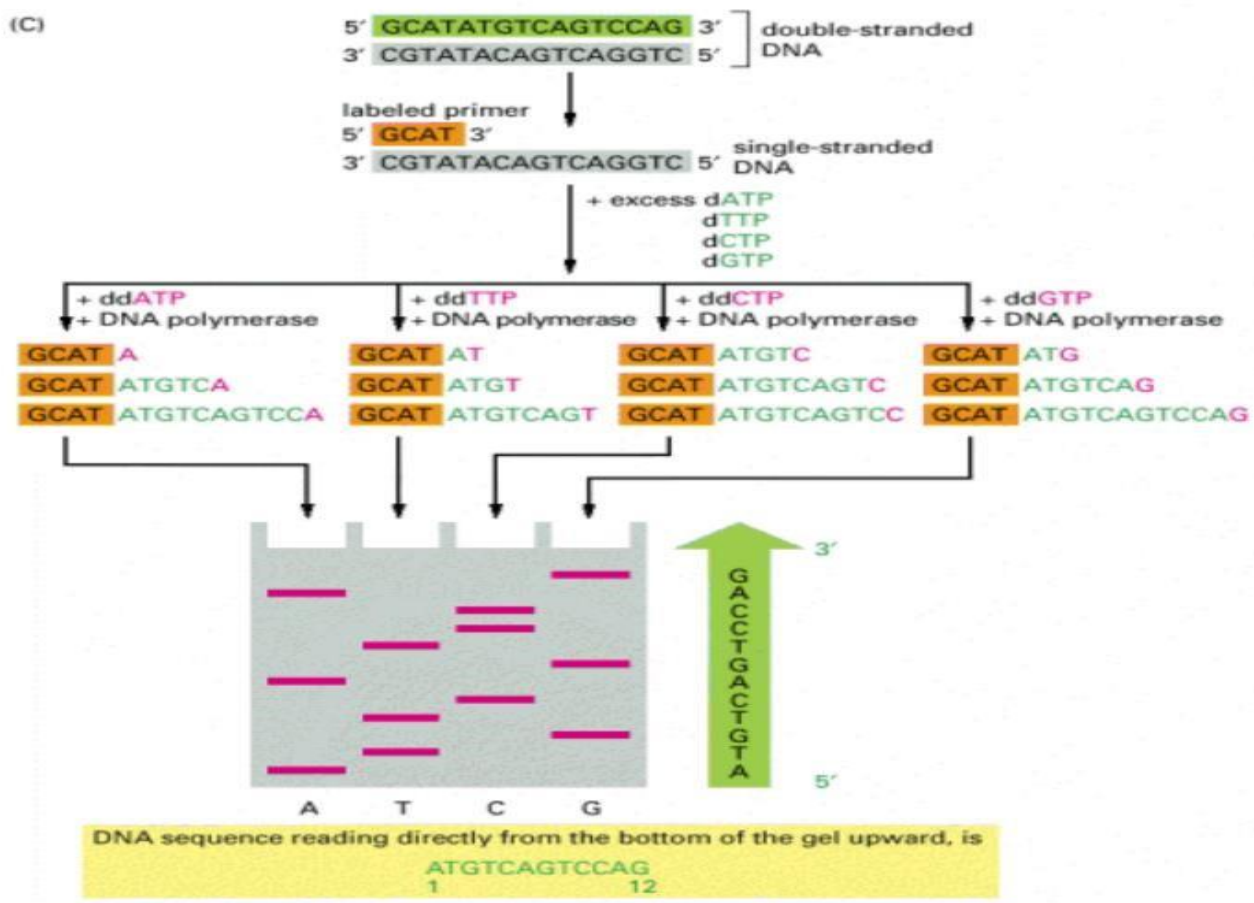
Now during the addition process, a dideoxynucleotides may be added to the first missing nucleotide, another random one may get added to the second one, another random one during the process would be added to

the third one and so on. And as we all know, the gel electrophoresis show the arrangement of the strands based on size, and the key to knowing the sequence here is strands differing by only one nucleotide, and this last added nucleptide is colored allowing us to know what it is exactly (a t g c) so it would look something like this on the gel :



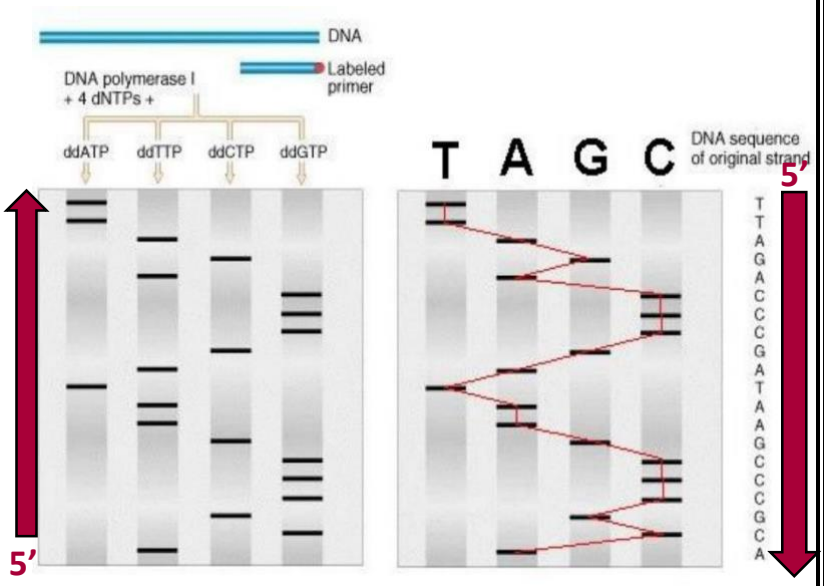
See how teach strand is different form the one before it by one colored nucleotide? They’re basically showing the sequence of the dna .

يعنيWhen we finally separate our sample, they will be separated according to size for sure, and in addition, due to the high resolution we will be able to distinguish between two fragments that only differ in one nucleotide, in other words, differing in one color.



Mechanism of determining the sequence post gel electrophoresis:

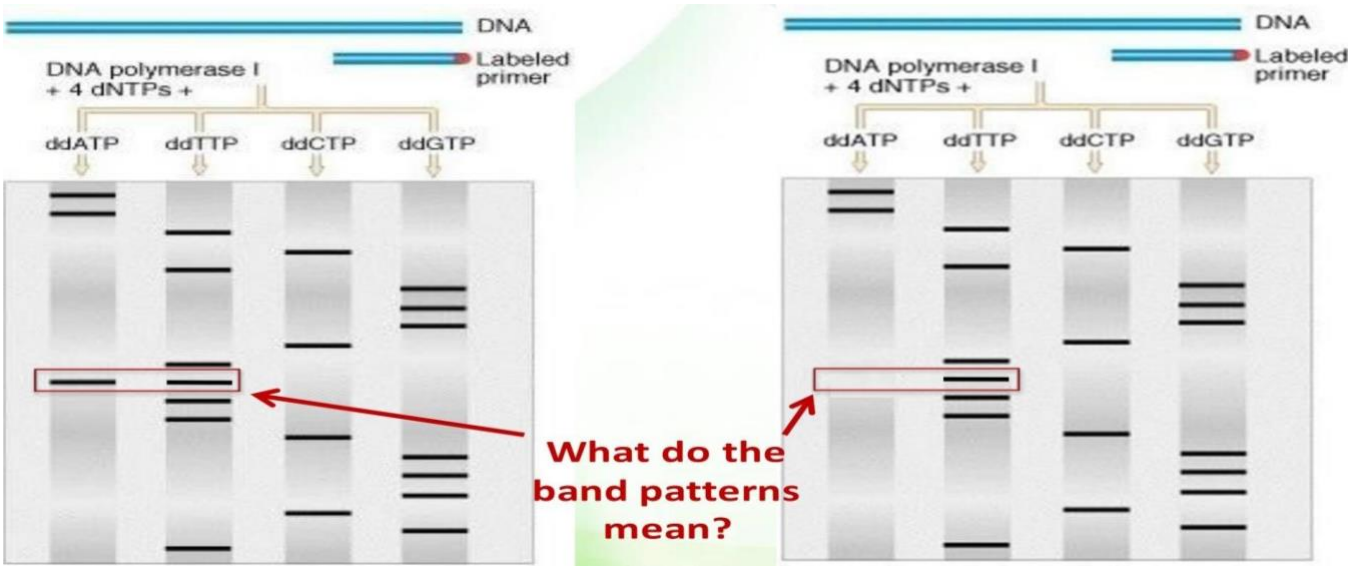
- Over here we have different results of gel electrophoresis, and from these results we can know the sequence of original **3'** DNA strand as it is illustrated in the picture, so how do we get these results? Simply we start from the last band(the shortest) this will be the first letter after the primer, here the shortest ends with T, but remember we



want the sequence of the original strand so it will be A in the original one, and we continue, the second shortest one ends with C so the original will have G, and remember we are reading from 5’ to 3’. 3’

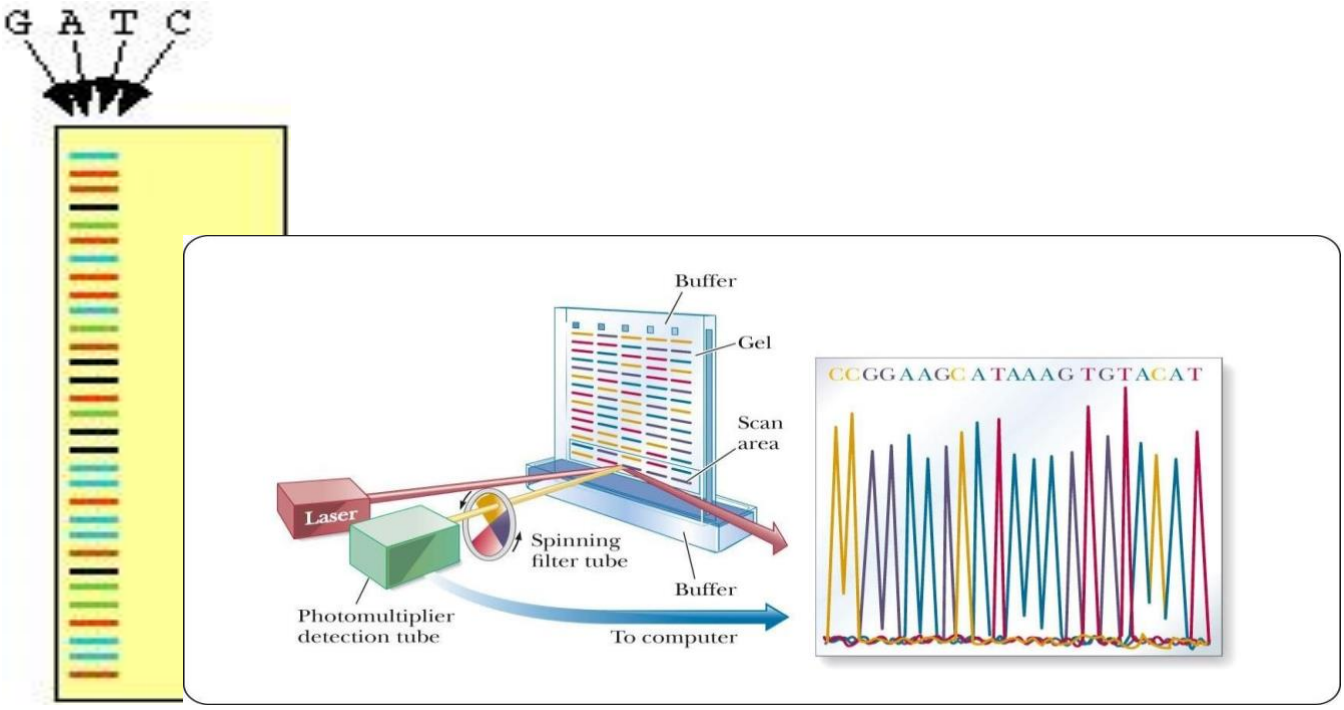
- Over here we have 2 bands that appear at the same level, so they have the same size, what does that mean?

❖ It means one of two things: either we have a mutation or polymorphism.

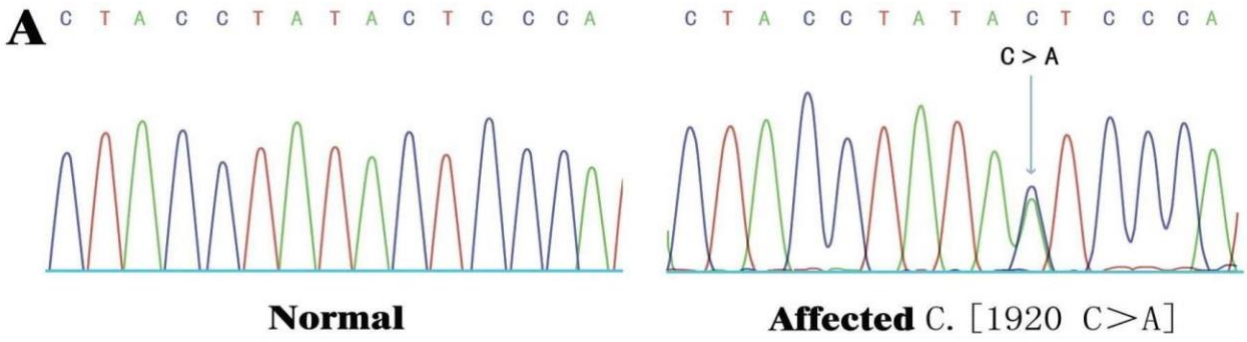


Heterozygous **Homozygous**

❖ Our cells are diploid so we have 2 chromosomes, so whenever we take DNA from an individual and we sequence that DNA, we are sequencing both chromosomes (the maternal and the paternal). And because of genetic variants, there is differences in DNA sequences among individuals (you can have A on one chromosome but G on the other chromosome), which can be considered as polymorphism (If it exists in more than 1% of the population) or mutation (If it exists in less than 1% of the population). And by looking at the figure above, you can see that the person on the right is homozygous; because there is only one band, but the one on the left is heterozygous; because we have two bands, and there might be a mutation on one of the chromosomes.



We will have large molecules at the beginning of the lane and the smaller ones further, and each one of them will give a certain color, so we will be benefited of this using sensors that can read the florescent tag, and the results will be displayed in the form of peaks (notice the waves in the figure), and each peak represents a certain letter (nucleotide), and that's what sequencing mean!



Using peaks, we can determine if the individual is normal or abnormal, as you can notice in the figure, the affected individual has two peaks in a certain location, which means that he is heterozygous (having 2 bands that overlap and they give signal), so the instrument will read two peaks in the same position (might be an indicator for polymorphism or mutations), this mutation might be homo/heterozygous, if it is hetero,

there will be 2 peaks representing 2 different letters, and If it is homo, it will present one peak, but a different one from what expected.

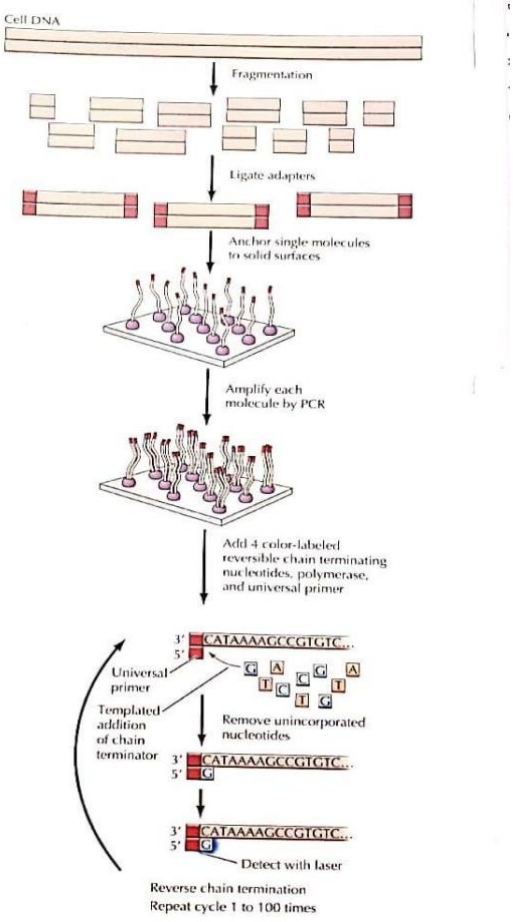
An overlap like this may not evfect the body at all, but may also have e biological effect.

So you might be:

1- Homo for the normal nucleotide (c,c)
2- Homo for the mutated nucleotide(A,A) 3- Hetero (A,C)

NEXT-GENERATION SEQUENCING

(THE FASTER AND THE BEST WAY) (IT ALMOST COSTS 100-500 DOLLARS)



- **Cellular DNA is fragmented. (randomly)**

Because we have millions of fragments, some of them will overlap.

- **DNA adapters (of known sequences) are added to ends of each DNA fragment. (same adapter for all of them)**
- **Each DNA fragment is attached to a solid surface and amplified like PCR using primers that anneal to the adapter sequences.**

We use the same primer for all fragments; because it is complementary with the adapter.

- **Four-color nucleotides with terminating ends are added.**
- **A single nucleotide is incorporated, and unincorporated nucleotides are removed.**

When a nucleotide is added, no other nucleotide is added, unless the added one becomes modified.

- **The incorporated nucleotide is modified for two reasons:**

➢ **It is activated and detected by a special camera. (and the color is translated into a letter)**

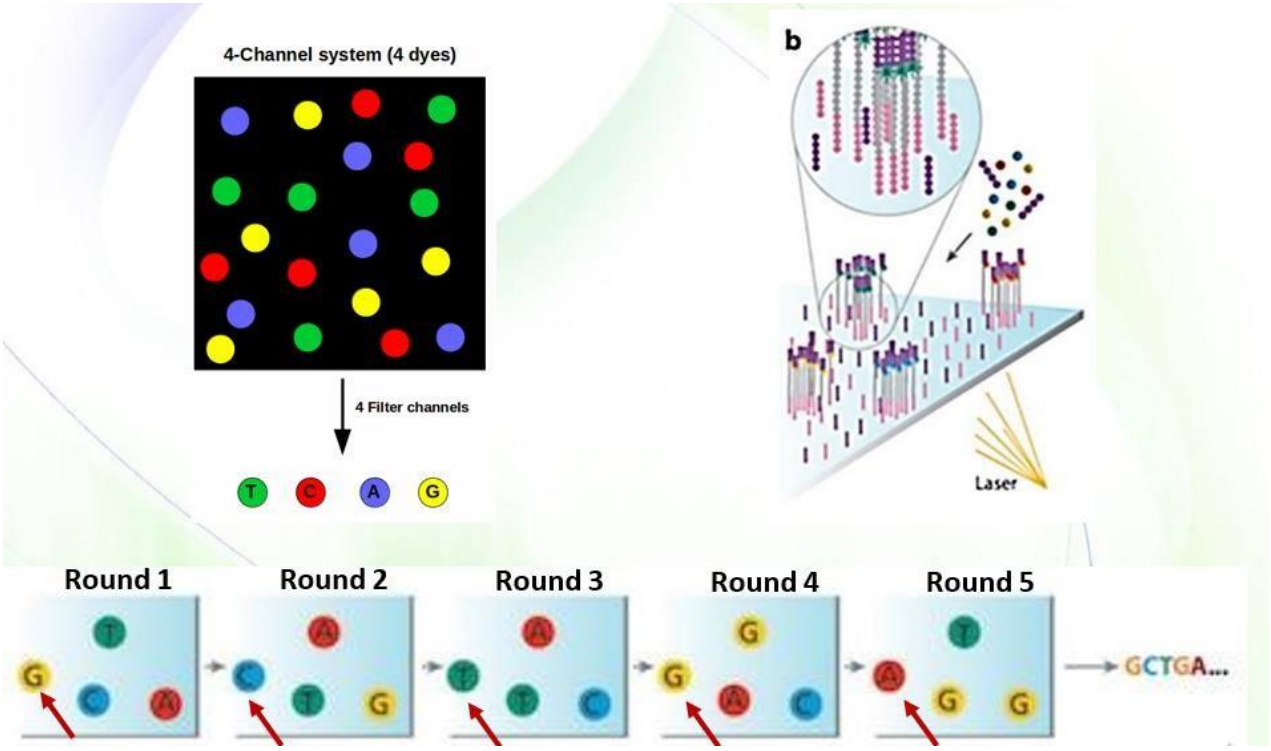
➢ **A new nucleotide can be added to it.**

 The cycle is repeated.

If you didn't understand I don't blame you.
what we are doing here is simply taking the DNA fragments and we put adapters of known sequences at each end of the fragment, we use the same adapter for all fragments, you will know why in a second.
we attach these fragments to a solid surface and we amplify them so we add primers that are complementary to the adapters(that's why we use the same adapter for all fragments; because we want to use a single primer and we want it to bind to all fragments) and the DNA fragments gets amplified by DNA polymerase -this enzymes uses nucleotides as substrates- so we add special nucleotides that have four different colors, one for each letter, and these nucleotides are special; they have terminating ends, meaning that when they are added no nucleotides are added after them, unless they get modified.

Soooooo, what we do is we let the DNA polymerase add the first nucleotide then we scan the colors, to know the first nucleotide that is added to each fragment, then we modify these nucleotides to allow the addition of another one, then we scan the colors again to know the second nucleotide that got added, and so on.

THE DETECTION



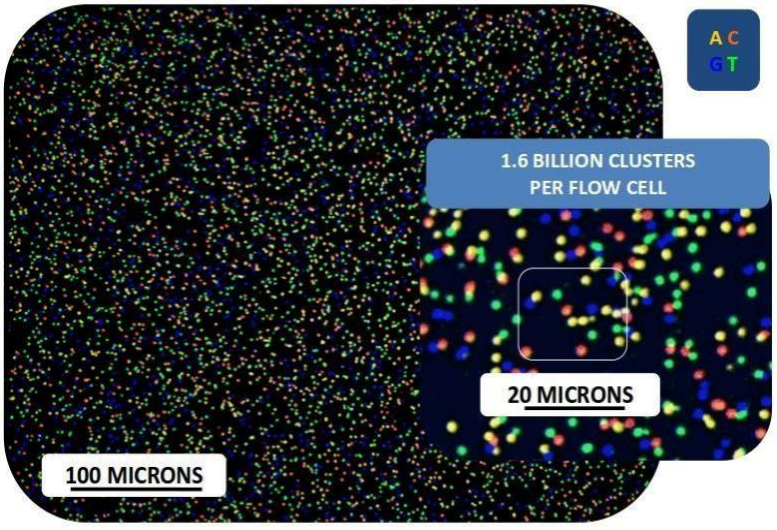
In a single round square above, each letter represents a whole cluster, so that means that we can sequence billions of fragments in the same time, and as proceeding in the rounds, we can get the full sequence, to make it more clear, the sequences above are: The top cluster: TAAGT

The bottom-right cluster: AGCCG

The bottom left cluster: CTTAG

The left cluster: GCTGA

When we are talking about billions of clusters, it will look like “fireworks”, the doctor said passionately, and all of them will be read at the same time.



Using bioinformatics, we would take the sequences of each structure, and since there is an

overlap, we can combine all the clusters in one sequence.
https://www.youtube.com/watch?v=womKfikWlxM

the doctor recommended you, precious colleagues, to watch this video about DNA sequencing, but he warns you from the amount of information inside it, so don't be worried about details, just understand the concept.

ضعيف القلب يخشى قلبه، وقوي القلب يخشى عقله.

تعلم فليسَ المرءُ يولدُ عالماً
صَغيرٌ إذا الْتَفَّتْ عَلَيهِ الْجَحَافِلُ
كَبيرٌ إذَا رُدَّتْ إليهِ المحَافِلُ

وَلَيْسَ أخو عِلْمٍ كَمَنْ هُوَ جَاهِلُ
وإنَّ كَبِير الْقَوْمِ لاَ علْمَ عِنْدَهُ
وإنَّ صَغيرَ القَومِ إنْ كانَ عَالِماً

جزيل الشكر و عظيم الامتنان لفريق بيوكيمسترز...

Good luck

Biochemasters