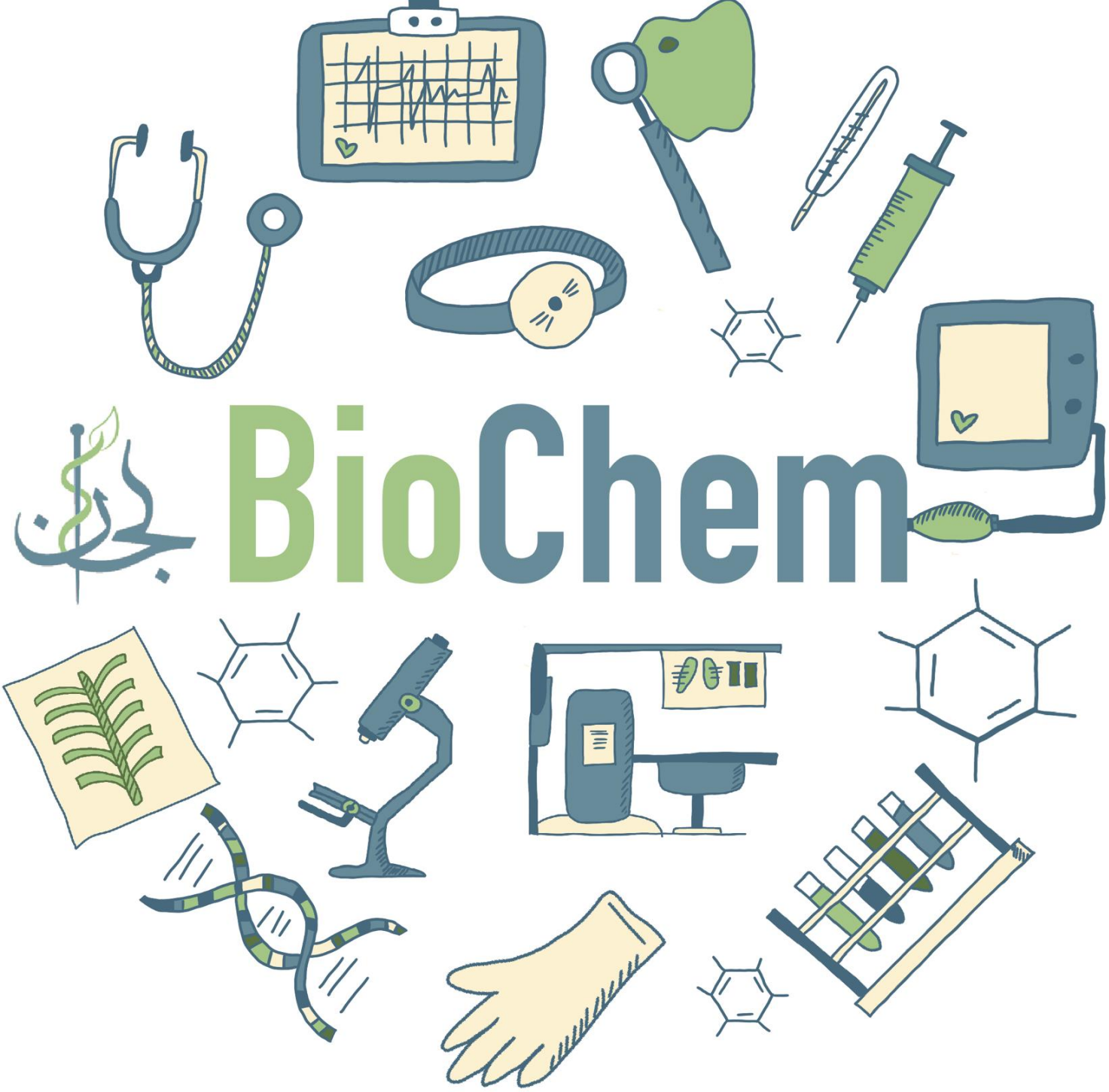


# BioChem


**Writer:** Hala Zaghloul & Ahmad Qatawneh

**Science:** Ahmad Qatawneh

**Grammar:** Ahmad Qatawneh

**Doctor:** Dr.Mamoun Ahram

# *Enzyme-based molecular techniques* ***(DNA sequencing)***

## What is DNA sequencing?

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

In other words, it's about knowing the order of nucleotides in a piece of DNA or a whole genome.

for example: you would know the sequence of a DNA segment from the starting point in that DNA segment ACGTAGCT etc... all the way to the end.

its almost like reading the letters of a book, you can identify those letters but its meaningless unless you know the code to translate and change these letters into meaningful words.

By knowing DNA sequence, it is possible to know many things: (**Importance**)

**1-Identification of genes and their localization**:

just like identifying words and their location in a book, it is possible to pinpoint a certain sequence that represents the beginning of a gene or the end of it, and also identifying where the introns and exons are.

**2-Identification of protein structure and function:**

identification of the amino acids sequence of a protein and predict the structure and functionality of this protein.

**3-Identification of DNA mutations:**

By knowing the normal sequence of a gene or a piece of DNA whether it is a coding region, promoter, or enhancer etc… it is possible to locate the abnormal sequences and identify mutations.

and by identifying the mutations, it is possible to tell if the mutation would cause the production of a defective protein that would lead to a certain disease or not.

**4-Genetic variations among individuals in health and disease:**

it is well established that humans are very much the same (99.9% of human DNA is the same for everyone with a a small room for variation 0.01% resulting from different phenomenon's including Single nucleotide polymorphism (SNP) . These sources of variation would have an effect on Human health and disease.

**5-Prediction of disease-susceptibility (قابلية) and treatment efficiency:**

knowing genetic variations in DNA molecules (throughout DNA sequencing) that affect the susceptibility of a person to develop a certain condition whether it was for instance , cancer or a heart disease.

**6-Evolutionary conservation among organisms:**

DNA sequencing is used in evolutionary sciences, that is the relationship and the connection of different organisms and to also learn about human migration throughout history.

## DNA sequencing of organism genome

Identification of DNA sequence started with smaller organisms like Viruses and prokaryotes first, then Human mitochondrial DNA was sequenced.

The first eukaryotic genome sequenced was that of yeast, Saccharomyces cerevisiae.

then the genome of a multicellular organism was sequenced, the nematode Caenorhabditis elegans.

Determination of the base sequence in the human genome was initiated in 1990 and completed in May 2006 via the **Human Genome Project**.

In 1990 the human genome project started; it was an initiative to sequence the whole human genome that's composed of 3 billion base pairs.

In 2006, they announced the completion of the sequencing of the human genome

*For your information: 3 billion dollars were spent on this project, 7 different countries got involved.

| SPECIES | BASE PAIRS (estimated) | GENES (estimated) | CHROMOSOMES |
|---|---|---|---|
| **Human** (*Homo sapiens*) | 3.2 billion | ~ 25,000 | 46 |
| **Mouse** (*Mus musculus*) | 2.6 billion | ~ 25,000 | 40 |
| **Fruit Fly** (*Drosophila melanogaster*) | 137 million | 13,000 | 8 |
| **Roundworm** (*Caenorhabditis elegans*) | 97 million | 19,000 | 12 |
| **Yeast** (*Saccharomyces cerevisia*) | 12.1 million | 6,000 | 32 |
| **Bacteria** (*Escherichia coli*) | 4.6 million | 3,200 | 1 |
| **Bacteria** (*H. influenzae*) | 1.8 million | 1,700 | 1 |

the picture above represents an overview of the differences in our genomes compared to different organisms, like mice and fruit fly.

you can see a little difference between the human genome and mouse genome, the number of genes is similar to each other though its not 25,000 anymore its about 25,500 now.

whereas in C.elegans the roundworm the number of genes is 19,000 which is also close to the number of genes in the human genome considering that we are talking about completely different species.

## The Size of Genomes (Nucleotides per genomes)

the size of genomes is also different, it's smaller in bacteria and archaea compared to humans.

It is possible to notice the diversity of protozoans, different plants and amphibians .they all have larger genomes than ours.

*you don't have to memorize all of these information.

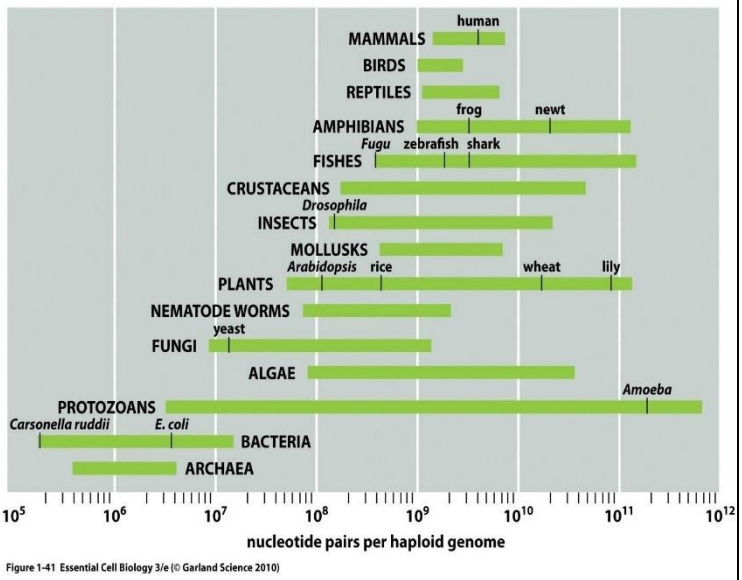
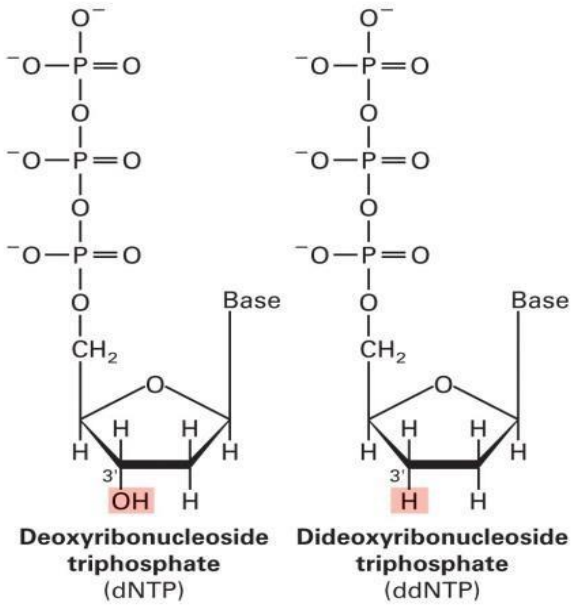


Figure 1-41 Essential Cell Biology 3/e (© Garland Science 2010)

## Method of DNA sequencing

The most popular method is based on premature termination of DNA synthesis (Also known as **Sanger-sequencing**) by dideoxynucleotides, so the basic technique was based on the use of a substrate known as ***dideoxyribonucleoside triphosphate (ddNTP***).




the normal substrate of DNA polymerase is **deoxyribonucleoside triphosphate (dNTP**), which has a deoxygenated carbon no.2 and the **hydroxyl group** on carbon no.3 the important one; the site where the 5' end of a nucleotide is added at in DNA synthesis which results in DNA elongation.

however, dideoxyribonucleoside triphosphate (ddNTP) has two deoxygenated groups on carbon no.2 and no.3. ( di = two)

which makes the 3' end deoxygenated, the absence of the hydroxyl group prevents the addition of another nucleotide, meaning that DNA synthesis would stop (**terminates**).

## The process

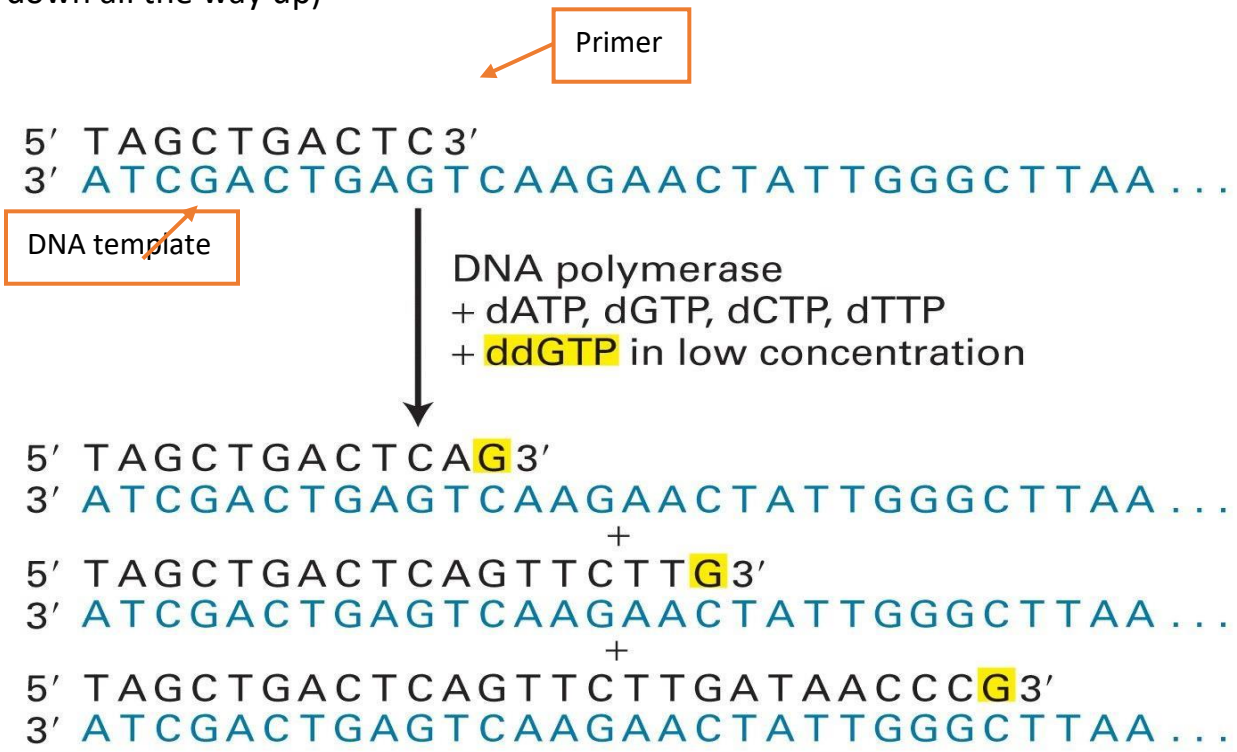
- DNA synthesis is initiated from a primer that has been labeled with a radioisotope for detection.
- Four separate reactions are run, each including deoxynucleotides plus one dideoxynucleotide (either A, C, G, or T).
- Incorporation of a dideoxynucleotide stops further DNA synthesis because no 3' hydroxyl 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 (from down all the way up)




In order to perform DNA-sequencing , we need to the following:

1- DNA template, the DNA to be sequenced.

2- DNA polymerase, to synthesize the DNA.

3- the 4 substrates dATP, dGTP, dCTP, dTTP.

4- **labeled** primer.

5- dideoxynucleotide.

- needing the previous components tells us that we should have some knowledge about the DNA fragment that's to be sequenced since the RNA primer must be complementary to a certain region of the DNA template.

- **four reactions** would take place, each one of them contains: the same template, primer, polymerase, substrates **but differ in having different dideoxynucleotides**.

so, the first reaction would have ddGTP (dideoxy-GTP), second one would have ddATP, the 3rd would have ddTTP and the $4^{th}$ would have ddCTP.

- when we talk about a DNA molecule, it doesn't necessarily mean we're talking about a single DNA molecule rather we're talking about thousands to millions of molecules of the same type.

meaning that everything is in abundance: DNA molecules, primers, substrates, molecules of DNA polymerase, which tells us that the substrates are NOT going to be all consumed.

**Conceptual Example:**

-let's say in a tube/vial there are 1,000 DNA template molecules (each would have a DNA primer attached to it), the DNA polymerase would start synthesis for each molecule.

as polymerase starts adding nucleotides it would either add deoxynucleotides OR dideoxynucleotides, referring to the picture in the previous page **ddGTP** or **dGTP** would be added to the newly synthesized DNA strand, but there is a greater chance that dGTP would be added because dideoxynucleotide is present at low concentration in the reaction, so ddGTP is at low concentration compared to that of D-GTP.

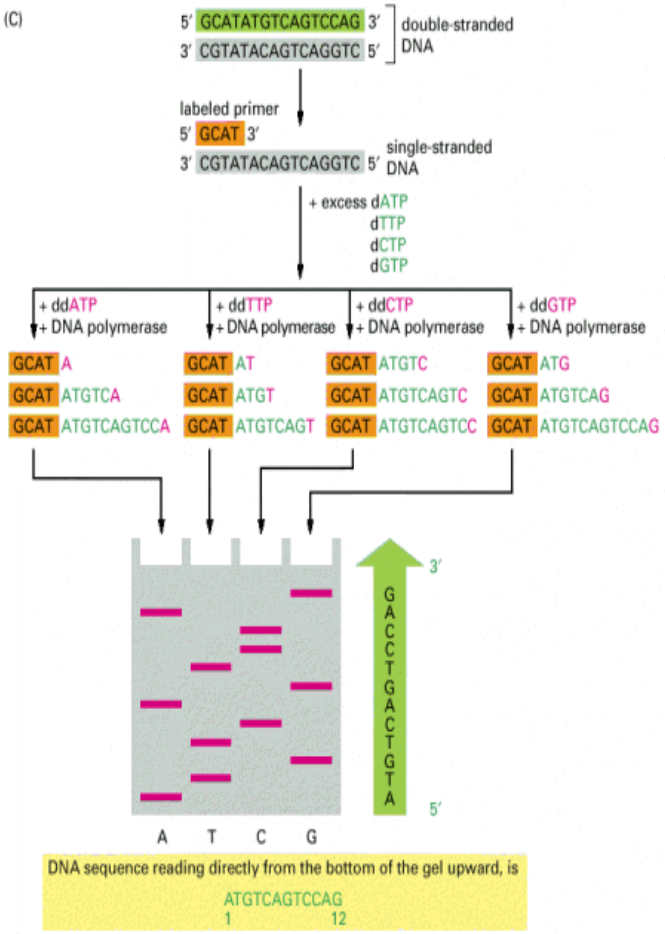
if dGTP is added DNA synthesis can continue, but if ddGTP is added DNA synthesis would be terminated and stopped.

-in this reaction 1,000 DNA template molecules are present, about 100 of them would be terminated and the other 900 can continue DNA synthesis, DNA polymerase will continue adding deoxynucleotides (dATP, dTTP, dCTP)

but when it comes to adding GTP again it would either add ddGTP or dGTP, this results in having another 100 DNA molecules with synthesis terminated, the other

800 DNA molecules would continue synthesis and another 100 DNA templates with synthesis terminated so on…

-this generates DNA fragments of different lengths.




this picture shows the different 4 reactions with different dideoxynucleotides.

in the reaction that has ddTTP, synthesis would be terminated at the sites where Adenosine nitrogenous base **(A)** is present in the DNA template, meaning in the **complementary newly synthesized DNA** fragment either dTTP or ddTTP would be added by DNA polymerase which results in DNA synthesis termination at different sites, generating fragments of different lengths .

-we take all of these reactions and put them in different wells, each well indicates that synthesis is terminated by a specific dideoxynucleotide.
e.g.: 1st well indicates the synthesis is terminated by ddATP, and so on…

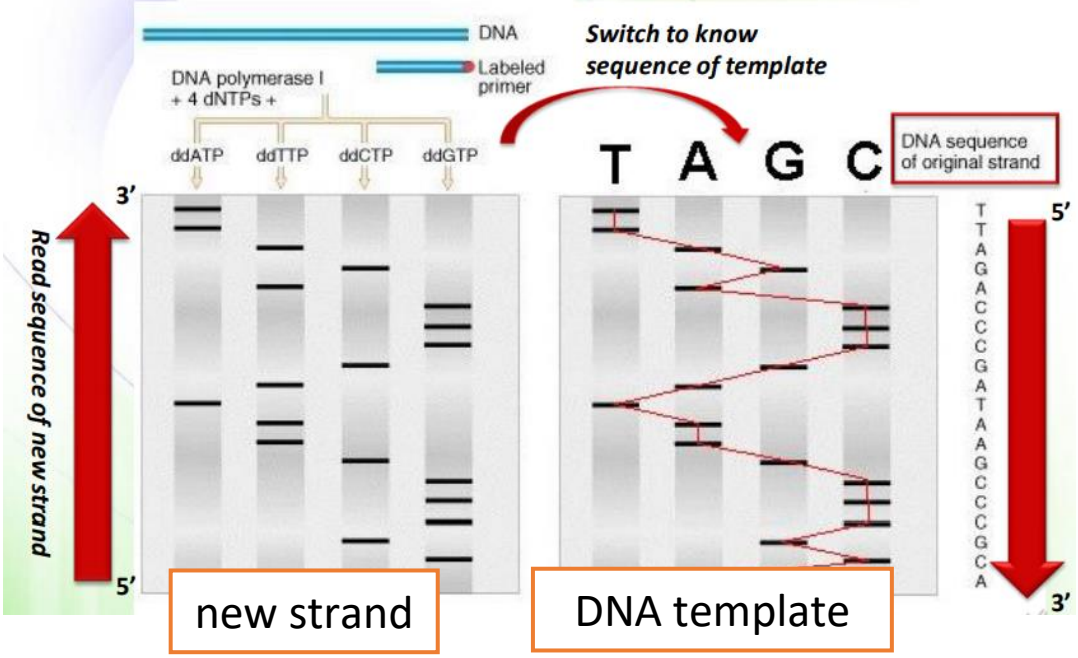
-smaller DNA fragments would migrate faster than the larger ones, their separation is based on one nucleotide, so based on the location of the fragment we can conclude the nucleotide the fragment has.

-each lane represents fragments of different lengths that have the same nucleotide, so fragments of the first lane have ddATP, 2nd lane fragments have ddTTP.

Using that information, It is now possible read the sequence of the newly synthesized DNA molecule; in which the smallest fragment ends with the nucleotide A, the one that's a little bit bigger ends with T, followed by G and so on. The sequence would surely be from 5' to 3' considering that the direction of DNA synthesis is from 5' to 3'.

(Notice the big green **arrow** in the picture above)

-




- Since Labeled Primers are used in these reaction , we were able to see the bands because there is signal (radiation) that is coming out of the Labelled fractionated DNA fragments, specifically from their primers.

on the left side of the picture the gel represents the fragments of the newly synthesized DNA and their sequence which is TGCGGGCTTATCGGGTCTAA.

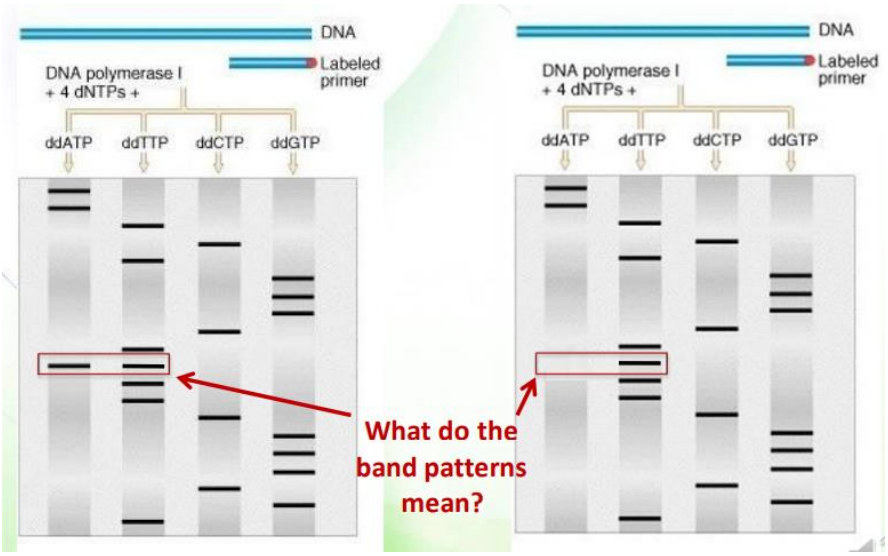
moving from the 5' end to the 3' end.

using that sequence, we can infer the sequence of the DNA template since its complementary to new synthesized strand.

the 5' end of the DNA template strand would start from the top part of the gel since the two strands are anti parallel to one another, and each lane would be complementary to the original one:

A→ T , T→ A , C→ G , G→ C]

- **what does it mean when there are two bands at the same exact position?**




## Explanation of the observation on the left :

- in the gel on the left there are two fragments at the same site, this means that they have the same exact length, this is possible because human cells are diploid, , that is they have two copies of each chromosome: one from the mother and one from the father. meaning that there is possibly a variation of one nucleotide between the mothers' DNA and the fathers' (**Heterozygosity**) DNA which happens to be the DNA sequenced in the previous image. so DNA polymerase is reading the same nucleotides from both DNA from the individual, but when there's a variation in one nucleotide, the mother's DNA is read **(T)** and the father's is read **(A)** for example, so both DNAs are read and both fragments would be detected on the gel at the same site because they have the same exact sequence differing only in one nucleotide also known as **Polymorphism** .
- Other possibilities can be seen including having a mutation in one of the alleles either mother's or father's, which results in the absence of a band on a certain lane on the gel, and instead it appears in another lane.

## Explanation of the observation on the Right:

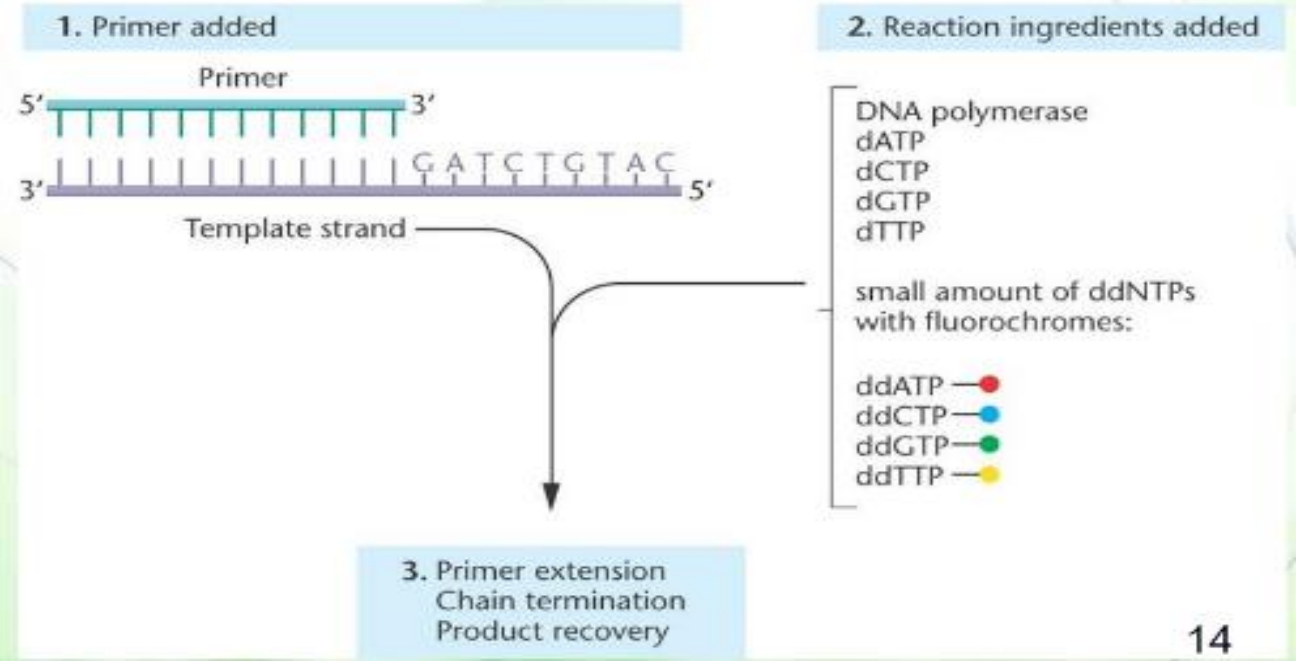
- It could be a mutation on both chromosomes (mothers and fathers) or maybe a variation in both chromosomes can cause changing in the nucleotides to ddTTP rather than ddATP for example (**Homozygosity**)

(Look at the right gel in the picture)

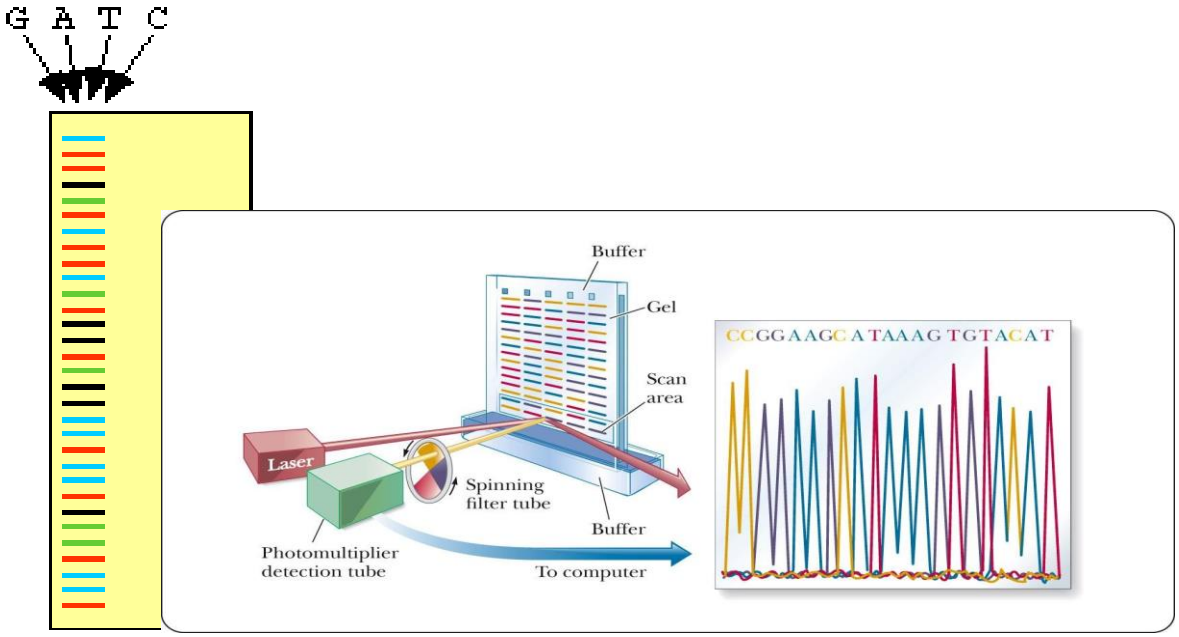
## *Fluorescence-based DNA sequencing*

Reactions include the four deoxynucleotides plus the four dideoxynucleotides in the same reaction with ***each ddNTP labeled with a unique fluorescent tag***.

radioactive labeled primers are used in the DNA sequencing we have mentioned, since radioactivity is harmful for our cells & DNA as well as being expensive, laborious and takes a lot of time to do DNA sequencing using radioactivity. scientists decided to use something less harmful like fluorescence which is sensitive and less dangerous. The cycle is repeated.



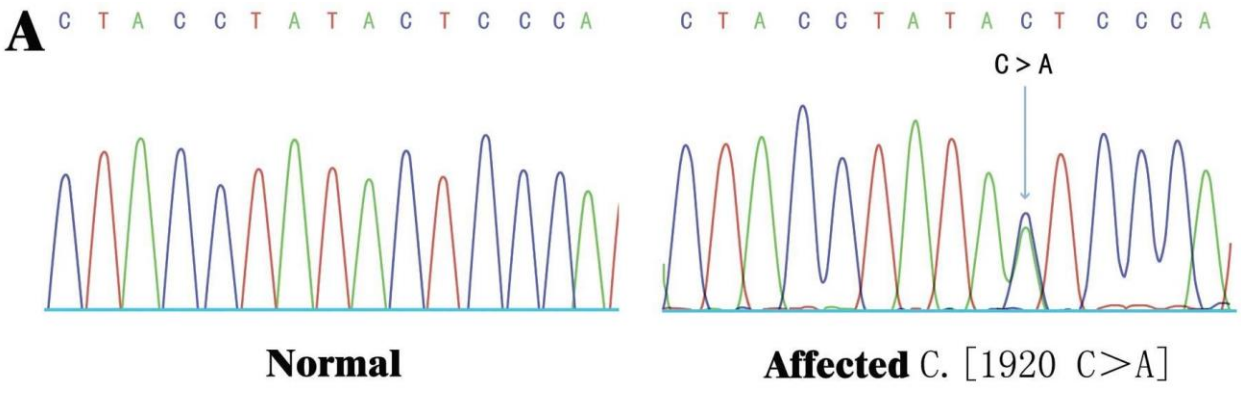

-the gel would be read by computers and the dideoxynucleotides would be labeled by fluorescence, instead of labeling primers, and each nucleotide would give a certain color throughout its own fluorescent signal.




-it's the same exact process, except that we use a single lane on gel electrophoresis since every type of the 4 types of ddNTPs emits a unique colored signal, and a detector would read and analyze the signal that would come out of each band, thereby identifying the DNA sequence since every florescent signal represents one type of a ddNTP. .

- the computer software would give us the signals in the form of peaks as shown in the picture, for example: the two yellow signals are read which means that the computer read C C dideoxynucleosides.

but **what if there's a variation or a mutation? how would it look like**?




## What does it mean?

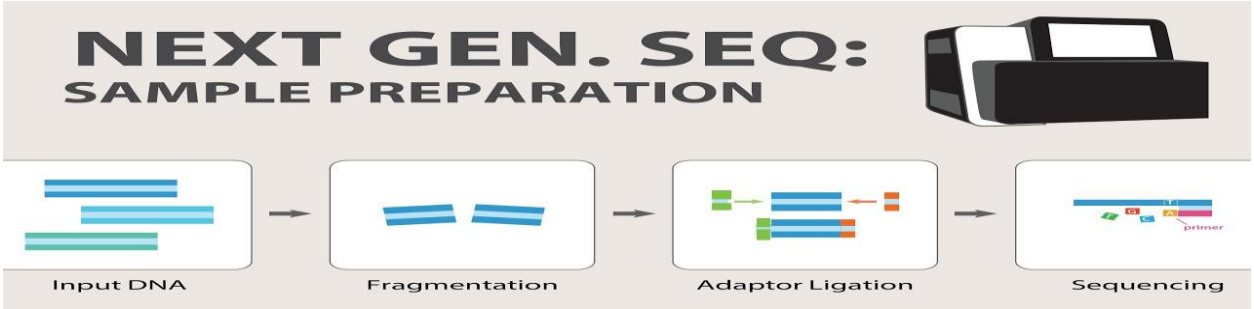
- Normally, we have peaks of different colors, but when two peaks overlap, the computer detects peaks of similar height, and interpret it as **polymorphism** which it means there is a **variation** between the mothers' and fathers' DNA just like we mentioned before. (page 10)

the signals would come from both the mothers' and fathers' DNA, if the mothers' DNA is read C and the fathers' G at a certain position **(Heterozygosity),** the DNA polymerase would read and add both nucleotides to the newly synthesized DNA fragments, some fragments would have G nucleotide, and others would have C nucleotide, but both portions are equally the same in terms of their lengths. which explains why the computer would read both dideoxynucleosides the same in the form of two peaks.

**what if there is a mutation on both alleles , for example instead of having a G on each allele, a C would be present instead ?**

- we would still have a single peak but representing a different signal than what we would **normally** exist.

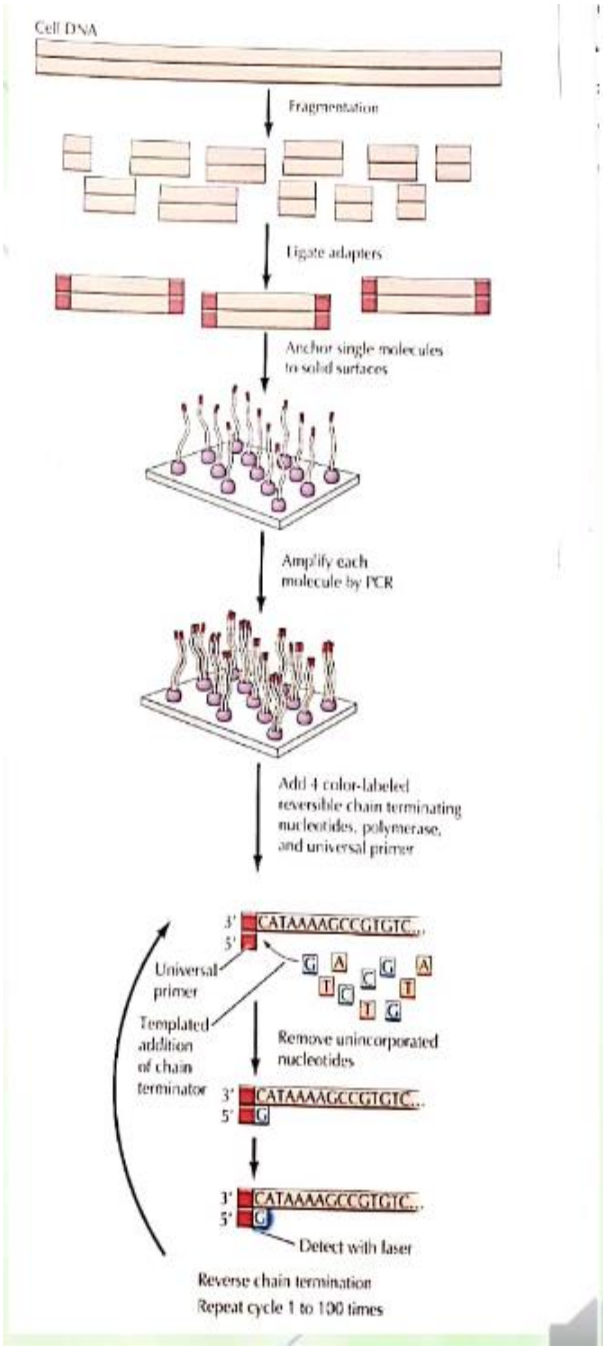
- Both techniques involving DNA-sequencing we have discussed previously are time-consuming and for that reason, there was a need for a efficient way to sequence DNA by utilizing Bioinformatics Known as **Next-generation sequencing.** What is Next-generation sequencing?



# *Next generation sequencing*

A group of scientists revolutionized DNA-sequencing by developing this technique in 2007 , including the scientist Greq-venter.

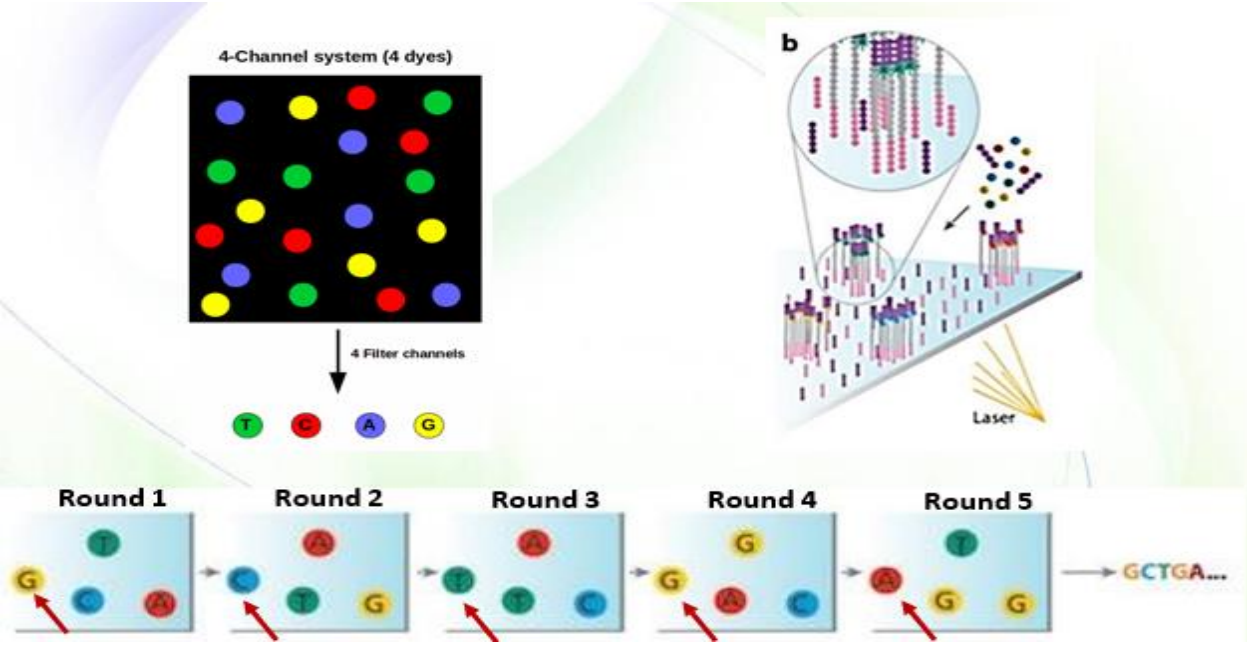
The principle of Next-generation sequencing: -




- Cellular DNA is fragmented randomly (notice that fragments are of different sizes)
- Identical DNA adapters are added **enzymatically** at the ends of each DNA fragment in order to :
    1- Allow DNA fragments binding a special platform (a solid surface)
    2- They can act as binding sites for primers
    3- DNA adapters can be used as tags to differentiate in case we are sequencing 2 different samples
- Each DNA fragment is attached to a solid surface and **amplified** (multiplication of each fragment) like PCR Using primers that anneal to the adapter sequences (Which are similar for all fragments)
- Four-color nucleotides with terminating ends are added . Note that these nucleotides are special nucleotides NOT deoxyribonucleotides that have the ability to fluoresce in 4 different colors and added substrate should be activated before the addition of a new nucleotide, then this nucleotide is activated (modified) by light rays : **(1)** releasing a fluorescent light (signal) which will be detected by a special camera **(2)** allowing the addition of a new nucleotide

- The cycle is repeated.

- Detection :-

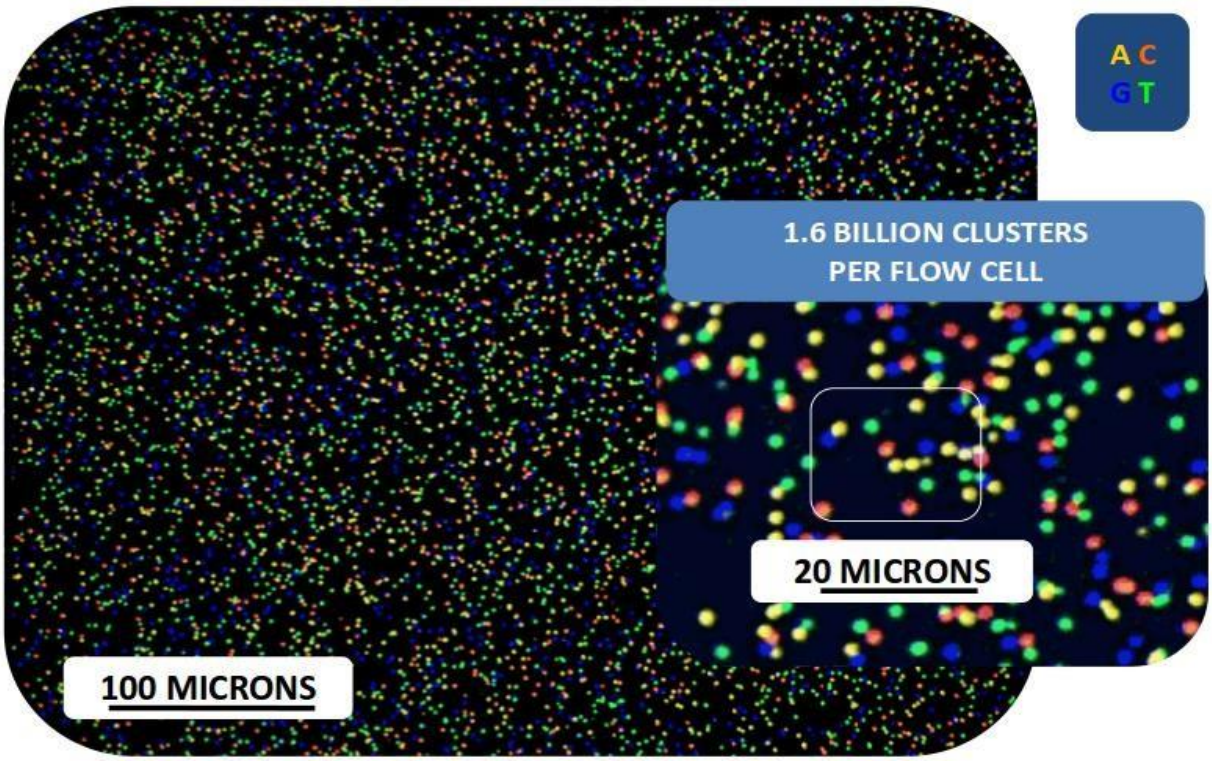



- We perform the previous steps in multiple rounds and in each round, a special camera-system records a fluorescent signal coming from the nucleotide until we sequence the entire fragments. (The entire human genome could be sequenced in this method within 24 hours).

Notes : 1- the signals are transmitted and detected within a small period of time .Furthermore , the florescent signal can only be transmitted provided that the Nucleotide is activated (It is excited by a light source)

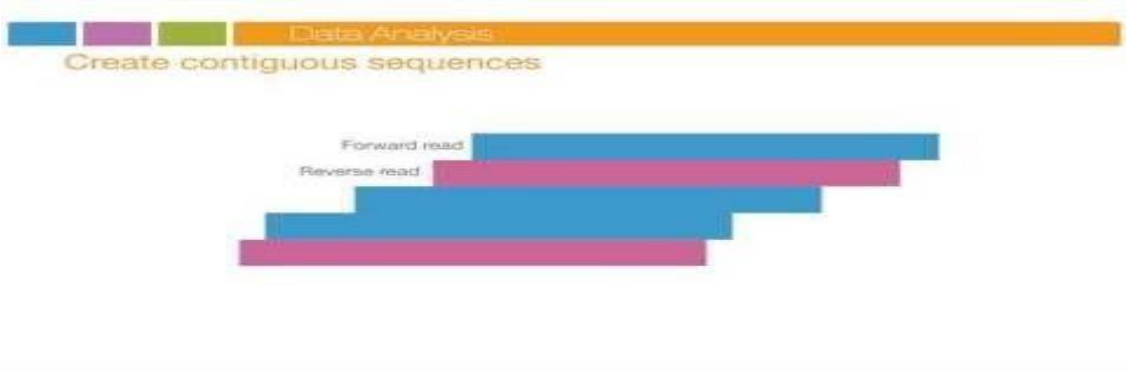
2-The addition of the subsequent nucleotide is very rapid because it is an enzymatically activated reaction and these reactions are known to be fast.

In the following Image, the camera system detects all of these signals in the same time and after that it organizes these signals in the form of a DNA-sequence.




**Final-Note** : the following video may contain extra details that are not required.

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




**The End**

# Self-assessment-Quiz

**Q1:-** Starting from the sequencing primer, what is the sequence of the original DNA sample ?

a) ATGACCGTA

b) TACTGGCAT

c) ATGCCAGTA

D) UACGGUCUA

e) TACGGTCAT


**Q2:-** 2. If we have 2 dATPs, 1 dCTP, 1 ddCTP, and 2 ddGTPs in one reaction tube, which of the following strands could be produced from a sample containing the following **template strand**: 5' GCTTGGCTTAACCAGATATTCCACTG 3' **with the following primer**: 5' CAGTGGAATATCTGGTT 3'?

a) 5' CAGTGGAATATCTGGTTAAG 3'

b) 5' CAGTGGAATATCTGGTTAAGCC 3'

c) 5' CAGTGGAATATCTGGTTAAGCCAA 3'

d) Just A and B

e) A , B and C.

**Q3:-** 5. If you wish to sequence a long strand of DNA in one round of reactions, you should:

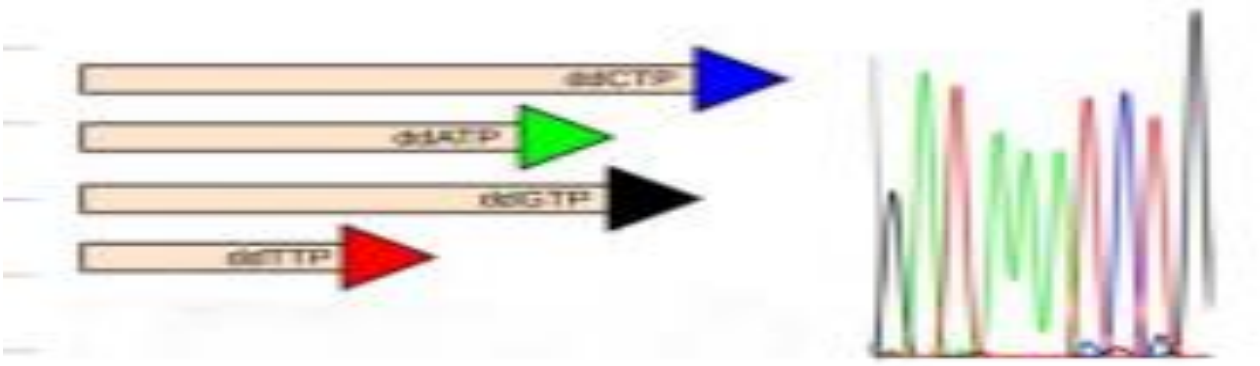
a) Increase the ddNTP/dNTP ratio

b) Decrease the ddNTP/dNTP ratio

c) Use a shorter DNA primer

d) None of the above

**Q4:-** What is the DNA sequence:




a) GTCTAAATAG

b) CTATTTAGAC

c) GATAAATCTG

**Answers:**

**1- B**
**2- A**
**3- B**
**4- C**