Note :

1. Currently we are facing issues installing Bioconductor packages in Kaggle . Hence we are unable to add code snippets. We will include code once the issue is resolved.

Context :

We are a team of Genomic Scientists , Bioinformaticians , Data Scientists , Clinical Scientists and Faculty scientists who are involved with the preparation of series of 15 tutorials in Genomic Data Science . The tutorials are aimed at non biologists to quickly grasp the essentials required to work in Genomic Data Science and also for existing biologists to brush up their knowledge when required.

1. Basics of DNA Analysis - https://www.kaggle.com/usharengaraju/basics-of-dna-analysis
2. Sequence Alignment - https://www.kaggle.com/usharengaraju/sequencealignment
3. Genomic Analysis - https://www.kaggle.com/usharengaraju/genomic-analysis
4. DNA Methylation - https://www.kaggle.com/usharengaraju/dna-methylation
5. RNA Sequencing - https://www.kaggle.com/usharengaraju/rna-sequencing
6. CHIP- Seq Analysis - https://www.kaggle.com/usharengaraju/chip-seq-analysis
7. Protein Analysis - https://www.kaggle.com/usharengaraju/protein-analysis
8. Proteomics - https://www.kaggle.com/usharengaraju/proteomics
9. Differential Expression Analysis - https://www.kaggle.com/usharengaraju/differential-gene-expression-analysis
10. Single Cell RNA Sequencing - https://www.kaggle.com/usharengaraju/single-cell-rna-sequencing
11. Phylogenetics - https://www.kaggle.com/usharengaraju/phylogenetics

Notebooks 12 to 15 -- Coming Soon

Basics of DNA and DNAAnalysis

Genomic Research is an important area of investigation into genomic make up of an organism. Organism could be unicellular(made of single cell) like Bacteria or multicellular (made up of many different cells) like any mammal. Anybody who is new to the area of studies would ask what is Genomic in the word Genomic Research? Every living organism has a set of genes (basic unit of heredity – basic layout to decide your physical, physiological traits or characters like skin colour, eye colour, type of hair, your shape of the face). The term Genome means complete set of genetic information in the form of genes present in a living system or organism. In other words, total gene information of a person. How do one person acquire this complete set of genes? Let's take human system, we have 23 pairs of chromosomes (one pair from mother, maternal and another pair from father, paternal). That why you see lot of resemblance between parents and their children.

Note : The kernel is not complete and its work in Progress . My co author for the kernel is Dr.Jyothirmayee.

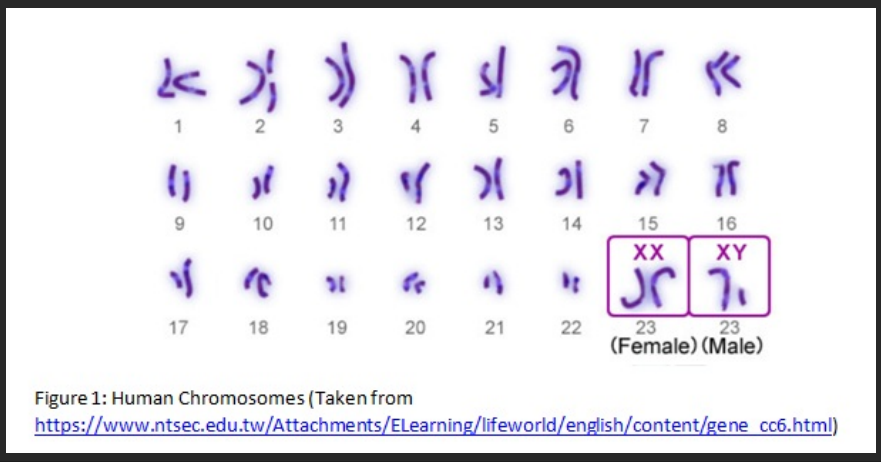


Figure 1: Human Chromosomes (Taken from https://www.ntsec.edu.tw/Attachments/ELearning/lifeworld/english/content/gene_cc6.html)

Next query would be, where does these chromosomes reside?

If you see, every cell has central dark nucleus and surrounded by cytoplasm with different components of a cell. These chromosomes (which carry genes) are present in nucleus in compact form. These genes have all the information needed for a cell to operate and survive the day to day battle with the environment.

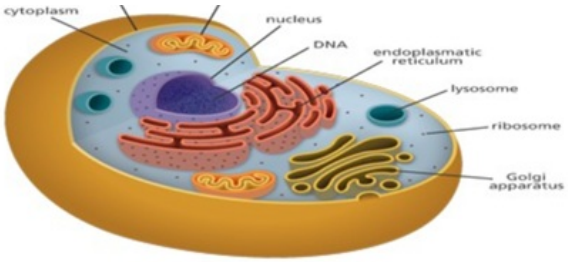
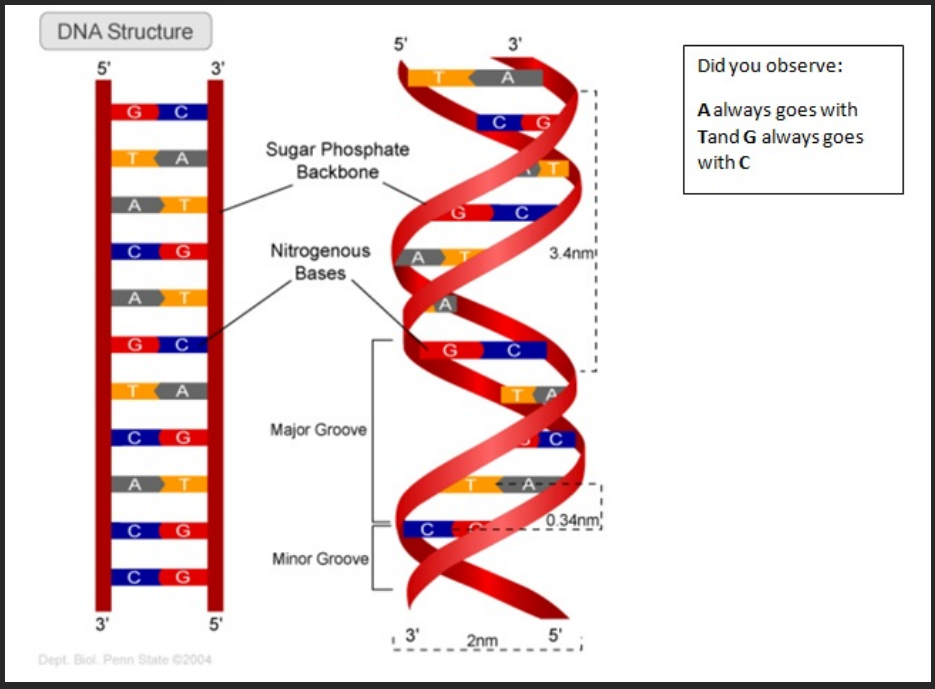


Figure 2: Human cell (Taken from https://www.assignmentpoint.com/science/biology/cell-structure-function-and-organisation.html)

Every cell of human body carries same information, but they behave differently depending where they exist. This is called differential gene activation based on the cell position and its function in the body. Let's take an example, heart cell versus liver cell. The basic character of heart cell is to show rhythmic contraction and liver cell is made to aid in food digestion…they are different in their function and character, yet they have same genomic information. Let's deep dive and understand what gene is made of off and how this genomic information residing in every cell nucleus is passed to their respective cells for them to work efficiently?



How would we represent a gene on a paper? A gene is nothing but part of DNA(Deoxyribonucleic acid) molecule. DNA or deoxyribonucleic acid is a long molecule that contains our unique genetic code.Genetic code means a language which every cell in our body understands. Genetic code is like a cypher code which holds the instructions for making all the proteins in our bodies. DNA is basically made up of four types of building blocks(nucleotides): A(Adenine), T(Thymine), G(Guanine) and Cytosine). Deoxyribonucleic acid is a molecule composed of two polynucleotide chains that coil around each other to form a double helix carrying genetic instructions for the development, functioning, growth and reproduction of all known organisms and many viruses.

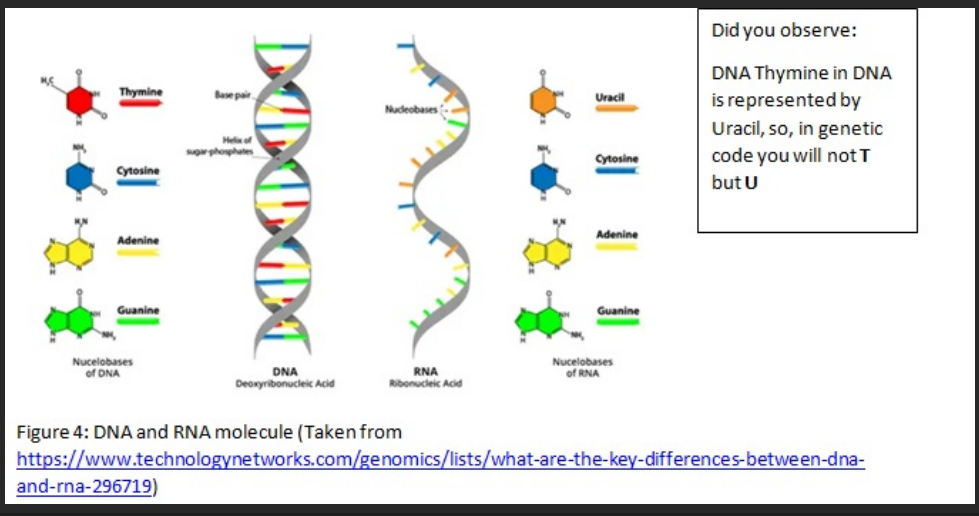
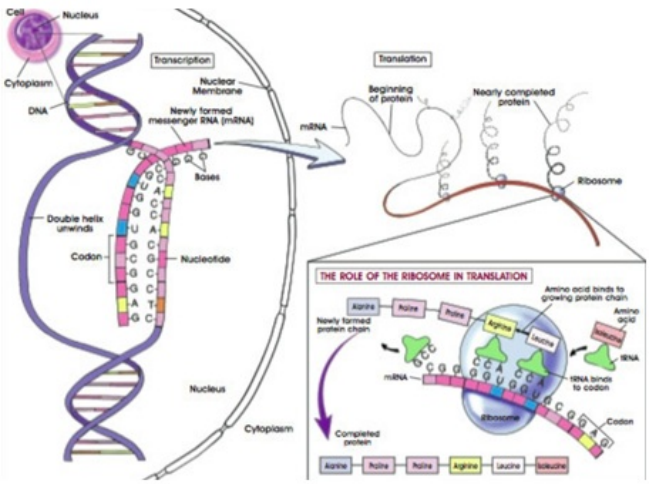


Figure 4: DNA and RNA molecule (Taken from https://www.technologynetworks.com/genomics/lists/what-are-the-key-differences-between-dna-and-rna-296719)

A gene can be easily represented as sequence of nucleotides – different variations and permutation combination of As, Ts, Gs & Cs. This information preserved in the nucleus of the cell mustbe conveyed to cytoplasm of a cell for it to produce proteins for fulfilling all the actions our body cells carries out. Let's assume a certain gene is switched on in the cell nucleus as the cell need to

express this gene for it to function. The DNA double helix unwinds and the information in that gene is copied to a different genomic molecule called RNA(Ribonucleic acid). RNA is a polymeric molecule essential in various biological roles in coding, decoding, regulation and expression of genes. This RNA is named as messenger RNA or in short mRNA. This takes the gene information from DNA and travels from central nucleus to outer cytoplasm and interacts with protein making cytoplasmic machinery called Ribosomes.



Did you observe:

Ribosome reads mRNA in sets of three nucleotides, called triplets

Figure 5: Protein Synthesis Process (Taken from https://socratic.org/questions/how-would-you-describe-the-steps-in-protein-synthesis)

Ribosomes are the protein builders or the protein synthesizers of the cell. This Ribosome binds with mRNA and based on the instruction in mRNA it organizes amino acids(building blocks of Protein) according to the sequence provided by mRNA and with the help of tRNA(transfer RNA: which gathers different amino acids from the surrounds of the cell), the cell makes a protein. So Information in nucleic DNA in the form of As, Ts, Cs, and Gs are transferred from nucleus to ribosomes present in cytoplasm and from there based on the DNA sequence, sequence of amino acids are put together and protein molecules are produced for it to carry out normal cell functions. Codon means sets of three nucleotide making up a amino acid molecule. So, if you have a Total sequence of all genes for a person, you can interpret many of the aspect of that persons like if that person carries a genetic defect like predisposition to diabetes or certain genes which may mutate and form cancer or a person's ability to with stand extraordinarily under extreme conditions. Genes combined with the environment cues forms the basic recipe of an organism or a living systemperformance. It define its physical traits, its metabolic conditioning, intellectual development (IQ, EQ) and its ability to survive(survival of fittest). There are 20 different amino acids.

The Genetic Code

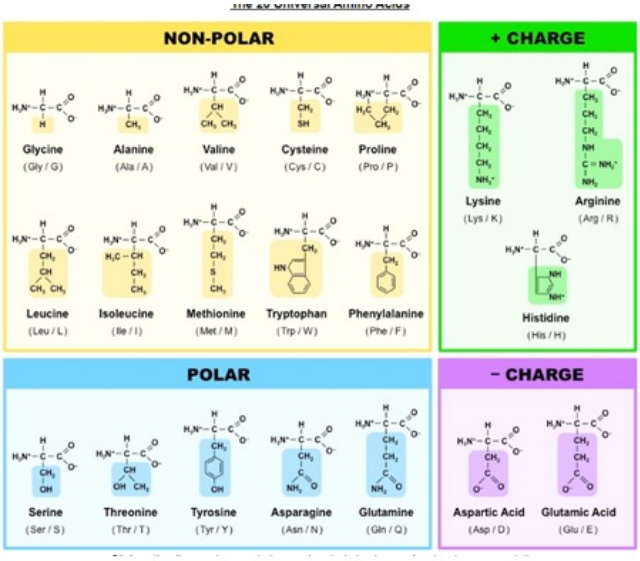
How do 64 different codons produce 20 different amino acids?

1st letter	Second Letter: U				Second Letter: C				Second Letter: A				Second Letter: G				3rd letter
U	UUU	Phe			UCU	Ser			UAU	Tyr			UGU	Cys			U
	UUC				UCC				UAC				UGC				C
	UUA	Leu			UCA				UAA	Stop			UGA	Stop			A
	UUG				UCG				UAG	Stop			UGG	Trp			G
C	CUU	Leu			CCU	Pro			CAU	His			CGU	Arg			U
	CUC				CCC				CAC				CGC				C
	CUA				CCA				CAA	Gln			CGA				A
	CUG				CCG				CAG				CGG				G
A	AUU	Ile			ACU	Thr			AAU	Asn			AGU	Ser			U
	AUC				ACC				AAC				AGC				C
	AUA				ACA				AAA	Lys			AGA	Arg			A
	AUG	Met			ACG				AAG				AGG				G
G	GUU	Val			GCU	Ala			GAU	Asp			GGU	Gly			U
	GUC				GCC				GAC				GGC				C
	GUA				GCA				GAA	Glu			GGA				A
	GUG				GCG				GAG				GGG				G

• The start codon is AUG. Methionine is the only amino acid specified by just one codon, AUG. • The stop codons are UAA, UAG, and UGA. They encode no amino acid. The ribosome pauses and falls off the mRNA. • The stretch of codons between AUG and a stop codon is called an open reading frame (ORF). Computer analysis of DNA sequence can predict the existence of genes based on ORFs. • Other amino acids are specified by more than one codon--usually differing at only the third position.

Taken from http://biology.kenyon.edu/courses/biol114/Chap05/Chapter05.html

The 20 Universal Amino Acids



Do you know, based on amino acid nature proteins folds. Any misfolding of Protein leads to disease like cataract in humans

Figure 6: Twenty Amino acids (Taken from https://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/24-proteins/amino-acids.html)

Genomic Research is gathering DNA or RNA or Protein sequence (genetic information) for different organisms and studying the pattern of gene occurrence or they distortion happening due to mutation or error in information transfer from cell to cell or from one generation to next generation and deriving useful data to guide the health of a person. Your next enquiry would be:how is this gene information gathered? Answer is by doing DNA or RNA or Protein sequence in a machine called Sequencer.DNA sequencing is the process of determining the sequence of nucleotides (As, Ts, Cs, and Gs) in a piece of DNA. Whole Genome sequencing of Human DNA was done in 2003 under a hugely popular project called Human Genome Project with collaboration of different countries, laboratories and genomic researchers with big bucks: a very expensive experiment. This paved way to, many such projects and at smaller budget as the technique got simpler, cheaper and easier to replicate in any standard lab. This involved making smaller bits of the whole genome, sequencing these small bits and combining the sequences based on a guiding or reference genome by Sanger sequencing method. What is this method? This method is a reverse engineering of the process happening in a cell. Sanger sequencing involves making many copies of a target DNA region. The raw ingredients are like those needed for DNA replication in an organism, or for polymerase chain reaction (PCR), which copies DNA in vitro. They include:

• A DNA polymerase enzyme for DNA synthesis or combining nucleotides (like stitching of the different nucleotides into a chain of nucleotide based on the instruction present on the bit of DNA as a guide

• A primer, which is a short piece of single-stranded DNA that binds to the template DNA and acts as a "starter" for the polymerase

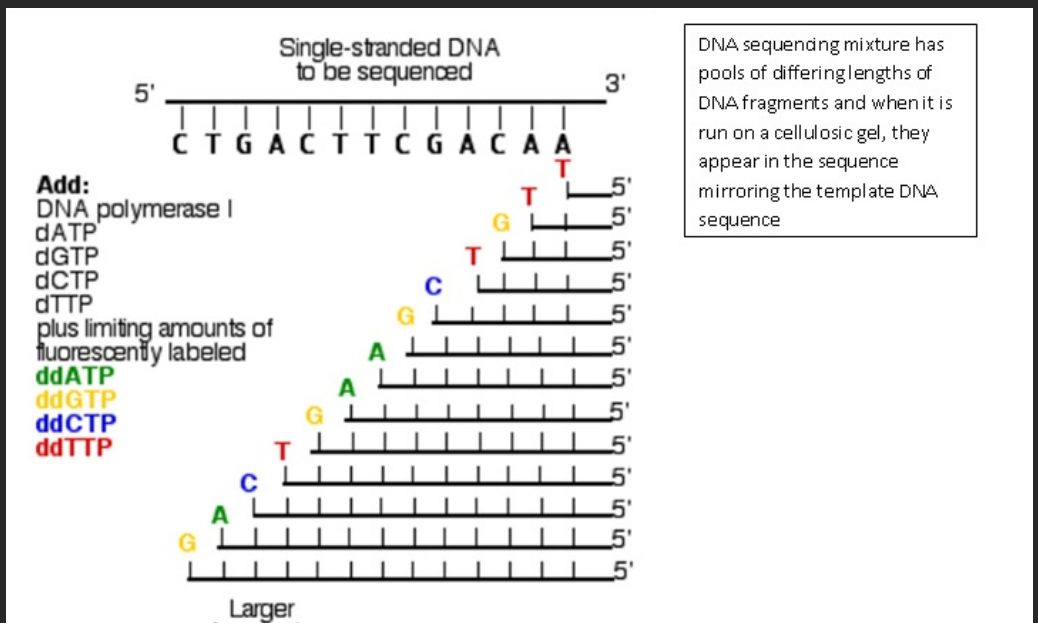
• The four DNA nucleotides (dATP, dTTP, dCTP, dGTP)- this a special nucleotide that fluoresces in different colours.Dideoxy, or chain-terminating, versions(ddATP, ddTTP, ddCTP, ddGTP)- dye attached nucleotides

• The template DNA to be sequenced

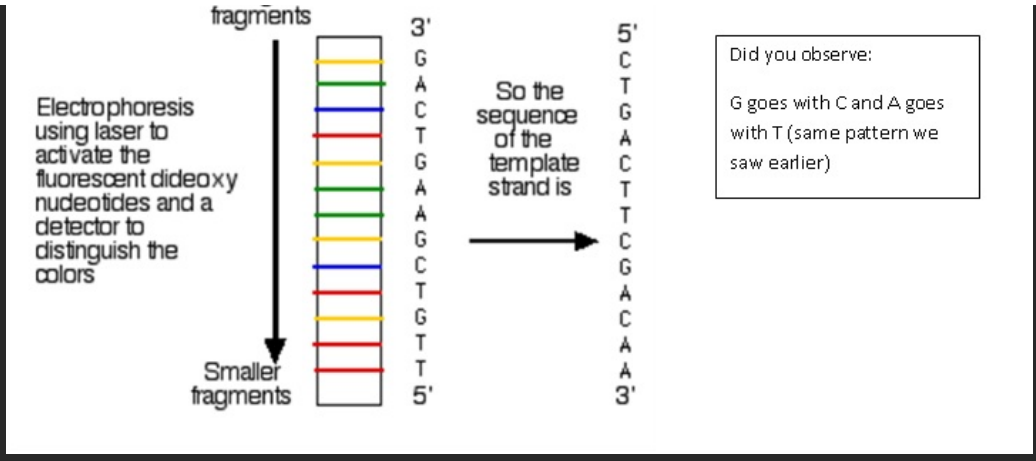
The whole process is carried out in a system which in the form thin gel and named process as capillary gel electrophoresis.

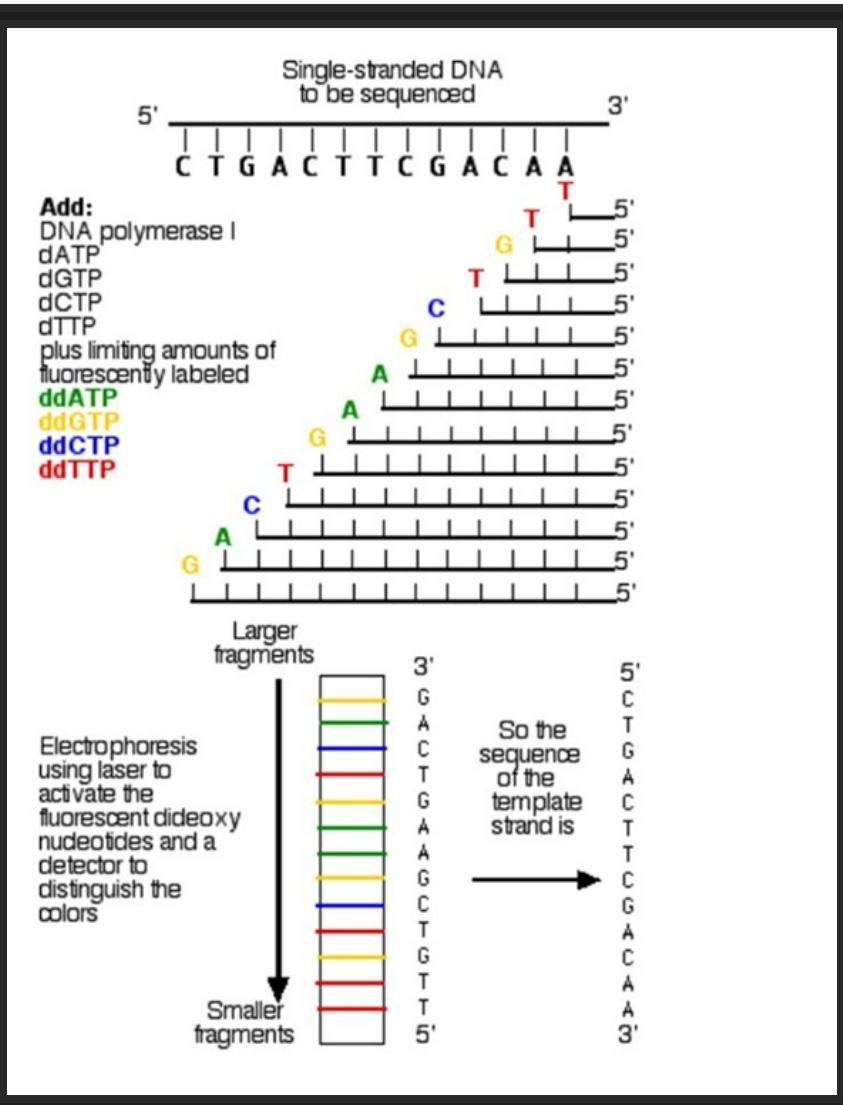
A mixture all different components are added and loaded on the gel, based of position of ATGCs, ddATP, ddTTP, ddCTP&ddGTPs starts joining a new sequence based on the template or guiding DNA.

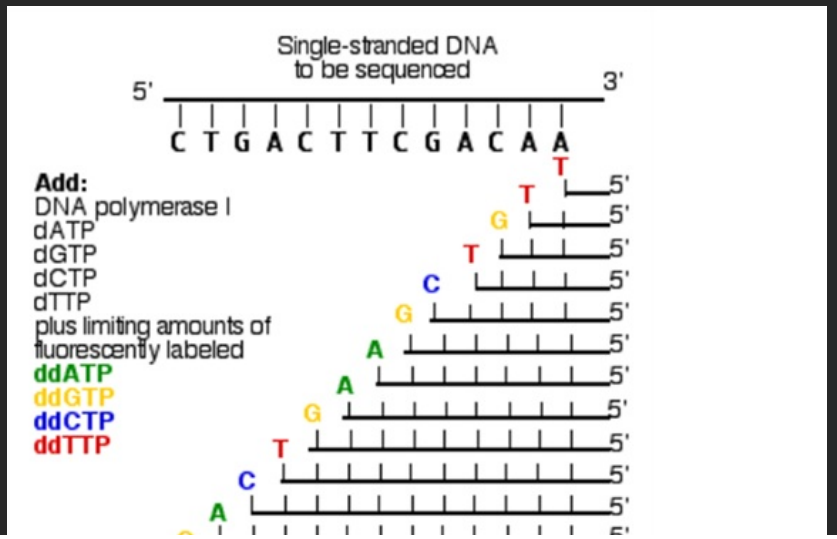
You will see a graphical representation below:



DNA sequencing mixture has pools of differing lengths of DNA fragments and when it is run on a cellulosic gel, they appear in the sequence mirroring the template DNA sequence


fragments
3'
G A C T G A A G C T G T T 5'
Electrophoresis using laser to activate the fluorescent dideoxy nucleotides and a detector to distinguish the colors
So the sequence of the template strand is
5'
C T T G A C T T C G A C A A 3'
Smaller fragments
Did you observe:
G goes with C and A goes with T (same pattern we saw earlier)


Single-stranded DNA to be sequenced
5'
3'
C T G A C T T C G A C A A
Add:
DNA polymerase I
dATP
dGTP
dCTP
dTTP
plus limiting amounts of fluorescently labeled
ddATP
ddGTP
ddCTP
ddTTP
Larger fragments
3'
G A C T G A A G C T G T T 5'
Electrophoresis using laser to activate the fluorescent dideoxy nucleotides and a detector to distinguish the colors
So the sequence of the template strand is
5'
C T G A C T T C G A C A A 3'
Smaller fragments


Single-stranded DNA to be sequenced
5'
3'
C T G A C T T C G A C A A
Add:
DNA polymerase I
dATP
dGTP
dCTP
dTTP
plus limiting amounts of fluorescently labeled
ddATP
ddGTP
ddCTP
ddTTP

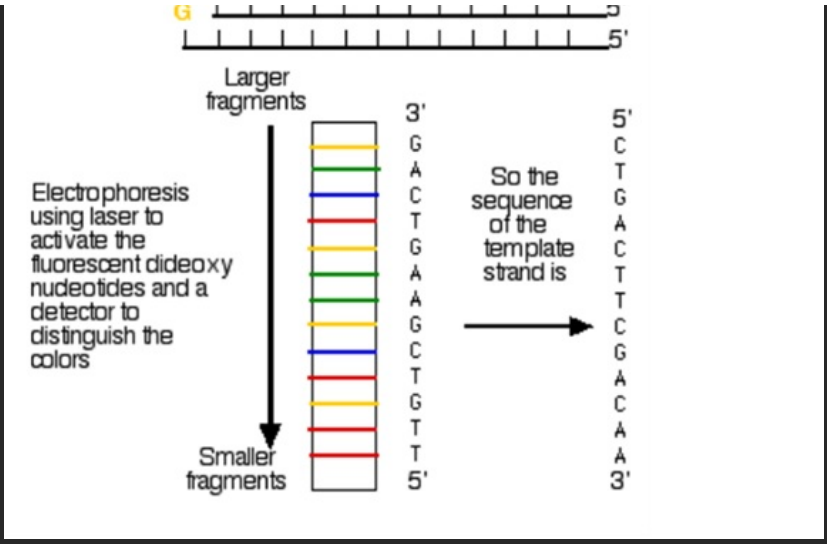


Figure 7: DNA sequencing using Sanger's Method (Taken from https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-biotechnology/a/dna-sequencing)

Next version of Sanger sequencing is Next Generation Sequencing with higher capability in speed and accuracy. The important features of Next Generation Sequencing are:

- Highly parallel: many sequencing reactions take place at the same time
- Micro scale: more like miniaturization of reactions: they are tiny, and many can be done at once on a chip
- Fast: because reactions are done in parallel, results are ready much faster
- Low-cost: sequencing a genome is cheaper than with Sanger sequencing
- Shorter length: reads typically range from 50 -700 nucleotides in length

Now, Large quantities of DNA can be sequenced much more quickly and cheaply with next-generation methods than with Sanger sequencing. For example, in 2001, the cost of sequencing a human genome was almost

$$100 million dollar. In 2015, it was just 1245.$$

With the advent of new technology, we can sequence different kinds of genomes and currently thousands of genomes have been wholly or partially sequenced.

With this we are flooded with genomic data in large volumes and the responsibility rests on data scientists and genomic researchers to dig deep for patterns and information which will help us fight or address different health issues.

All genomic data is stored in GENBank, a publicly available open source collection. https://www.ncbi.nlm.nih.gov/genbank/ GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data daily.

How are these sequences stored? It is stored in a standard format both in nucleotide and protein sequences. Nucleotide sequence and FASTA defline formats (.fsa):

- No size limit on nucleotide sequence, generally.
- FASTA file should consist of a single definition line beginning with a '>'.

Minimum requirements for the FASTA defline are:

- SeqID (sequence identifier) which is the text between the '>' and the first space. The SeqIDs limits are:
- Must be <50 characters
- Can only include letters, digits, hyphens (-), underscores (_), periods (.), colons (:), asterisks (*), and number signs (#).
- Organism and related information (unless organism information is included with -j at the command line or in a .src file)

Let's look at a sample for you get an idea:

Example FASTA

```
>Sc_16 [organism=Saccharomyces cerevisiae]
tataggcgaatcgagtatattatttttctcaacatatgtat
atgaacatgagaatatatttataggaatgtataaaattgtga
cctctcctgctattttagttactgattttatgtatgtagggg
gaataggggctgcctttcttaatgcagttttaatttttttctt
ttaatttttttcttagtaaaattatttaaagtaaagattaatg
```

```
gaataaccattgcgctttttttacagttttggtttttcat
tttttggaaaaaatattttaaatattttacctttttatttag
ggggtattttatatagtatctatacttcaacagatttttctg
aacatatagttcctattgctttttcaagtgcattagcccctt
ttgtaagcagtgttgctttttatggagaaatatcctatgaaa
catcatatataaatgcaattttaattggtattttaattggtt
ttatagtggttcctttgtctaaaagtctttatgactttcatg
agggatatgatttatataatttaggttttacagcaggtt
```

Protein sequence format (.pep)

• It is FASTA file of the protein sequence, where the SeqID must match protein_id in the .tbl file

Example FASTA

```
>WS1030 [gene=sde3p] [protein=SDE3P]
MYKIVTSPAILVTDFMYVGGIGAAFLNAVLIFSFNFFL
VKLFKVKINGITIAAFFTVFGFSFFGKNILNILPFYLG
GILYSIYTSTDFSEHIVPIAFSSALAPFVSSVAFYGEI
SYETSYINAILIGILIGFIVVPLSKSLYDFHEGYDLYN
LGFTAG
```

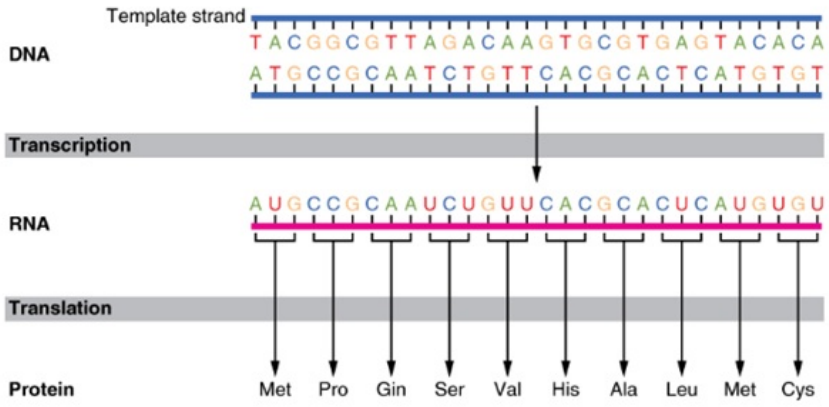


Fig 8: Process of information passage from DNA to RNA and finally to Protein (Taken from https://oerpub.github.io/epubjs-demo-book/content/m46032.xhtml)

This was a primer to get you acquainted with the basics of DNA analysis.

How do we start analysis of DNA sequences? It is done by comparing similarities and dissimilarities in sequences by alignment-based analysis like pairwise or multiple alignment (based on the number sequences compared). In next part of these article series we will discuss how to analysis DNA sequences which you can directly download from GenBank anytime and from anywhere.

In [ ]:
```
library(seqinr)
```

In [ ]:
```
corona <- read.fasta(file = "../input/customer-frauddata/coronavirus.fasta")
```

In [ ]:
```
coronaseq <- corona[[1]]
```

In [ ]:
```
length(coronaseq)
```

In [ ]:
```
table(coronaseq)
```

In [ ]:
```
GC(coronaseq)
```

In [ ]:
```
count(coronaseq, 1)
```

In [ ]:

```
count(coronaseq, 2)
```

In [ ]:
```
starts <- seq(1, length(coronaseq)-2000, by = 2000)
n <- length(starts)
chunkGCs <- numeric(n)
```

In [ ]:
```
for (i in 1:n) {
        chunk <- coronaseq[starts[i]:(starts[i]+1999)]
        chunkGC <- GC(chunk)
        print(chunkGC)
        chunkGCs[i] <- chunkGC
    }
```

In [ ]:
```
plot(starts,chunkGCs,type="b",xlab="Nucleotide start position",ylab="GC content")
```

In [ ]:

In [ ]:

References:

1. What is DNA?https://www.yourgenome.org/facts/what-is-dna
2. DNA Sequencinghttps://www.biology-pages.info/D/DNAsequencing.html
3. DNA sequencinghttps://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-biotechnology/a/dna-sequencing
4. Whole genome sequencinghttps://en.wikipedia.org/wiki/Whole_genome_sequencing
5. GenBank Overviewhttps://www.ncbi.nlm.nih.gov/genbank/
6. How would you describe the steps in protein synthesishttps://socratic.org/questions/how-would-you-describe-the-steps-in-protein-synthesis
7. DNA Sequencinghttps://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-biotechnology/a/dna-sequencing