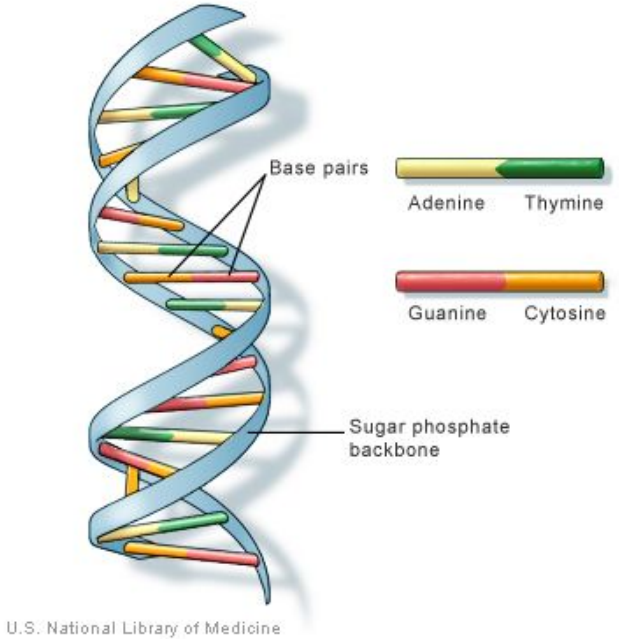
Dynamic Programming and Sequence Alignment

Application of LCS algorithm to DNA Analysis

What is DNA? What makes a DNA?

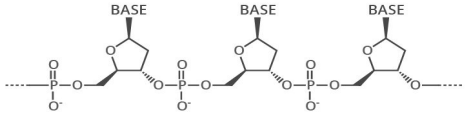
- DeoxyriboNucleic Acid.
- Molecule containing genetic code of Organisms.
- It is the basic building block of life of every kingdom.
- Nucleotides and Chromosomes make the DNA.
- Double Helix structure, which looks like a ladder twisted into spiral.
- Each step is a nucleotide pair.



THE CHEMICAL STRUCTURE OF DNA

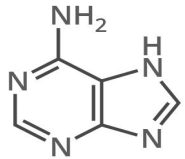
DNA (deoxyribonucleic acid) carries genetic information in all multicellular forms of life. It carries instructions for the creation of proteins, which carry out a wide range of roles in the body.

THE SUGAR PHOSPHATE 'BACKBONE'

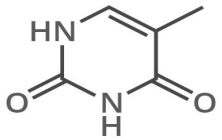


DNA is a polymer made up of units called nucleotides. The nucleotides are made of three different components: a sugar group, a phosphate group, and a base. There are four different bases: adenine, thymine, guanine & cytosine.

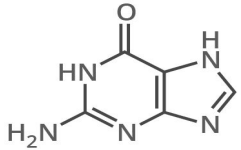
A ADENINE



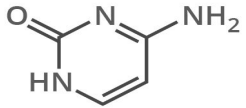
T THYMINE

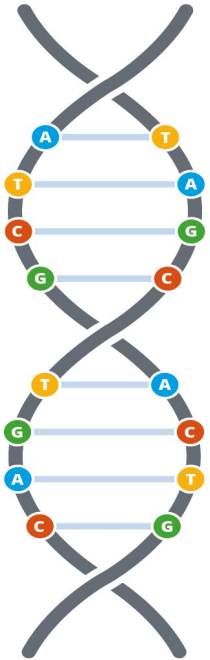


G GUANINE



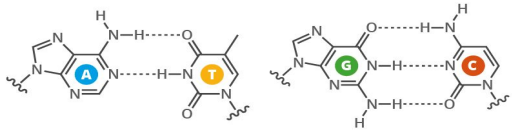
C CYTOSINE





WHAT HOLDS DNA STRANDS TOGETHER?

DNA strands are held together by hydrogen bonds between bases on adjacent strands. Adenine (A) always pairs with thymine (T), whilst guanine (G) always pairs with cytosine (C).



FROM DNA TO PROTEINS

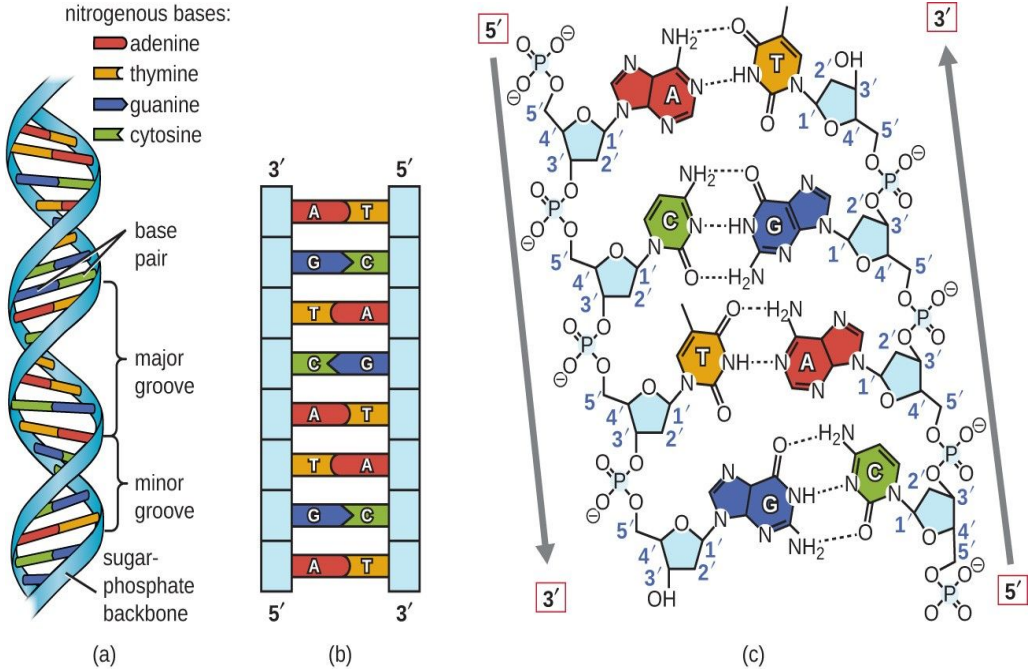
The bases along a single strand of DNA act as a code. The letters form three letter 'words', or codons, which code for different amino acids - the building blocks of proteins.

An enzyme, RNA polymerase, transcribes DNA into mRNA (messenger ribonucleic acid). It does this by splitting apart the two strands that form the double helix, then reading a strand and copying the sequence of nucleotides. The only difference between the RNA and the original DNA is that in the place of thymine (T), another base with a similar structure is used: uracil (U).

In multicellular organisms, the mRNA carries genetic code out of the nucleus, to the cell's cytoplasm. Here, protein synthesis takes place. 'Translation' is the process of converting turning the mRNA's 'code' into proteins. Molecules called ribosomes carry out this process, building up proteins from the amino acids coded for.

Why Understand DNA?

- Disease Diagnosis and Treatment
- Paternity and Legal Impact.
- Forensics
- Agriculture
- Pharmaceutical Production.
- Gene Therapy.
- Phylogenetics.



DNA is the future of computation.

Comparing DNA Strands

- Basis for understanding how different species differ from each other.
- understand its features, function, structure, or evolution
- gives information on the relationship between individual organisms, or between groups of organisms. It shows how closely related they are.
- A, C, G, T represent four nucleotide bases of a DNA strand - Adenine, Cytosine, Guanine, Thymine.

DNA Sequencing And Clustering

- Sequencing is the process of determining the precise order of nucleotides within a DNA molecule.
- Any technology to determine the order of the fours bases i.e. Adenine, Guanine, Cytosine and Thymine in the DNA strand.
- DNA Sequencer, which analyses light Signals from fluorochromes attached To the nucleotides.
- Human Genome stored on 23 chromosome Pairs in nucleus and mitochondria.
- Clustering is the process of grouping the DNA sequences that are related.

Where does DNA Sequencing, comparison and clustering come into picture?

- Knowledge of DNA sequences has become indispensable for basic biological research and DNA sequencing and clustering are methods to process a DNA sequence which help determine unique genes existence
- These methods are used in numerous applied fields such as medical diagnosis, biotechnology, forensic biology, virology and biological systematics.
- The knowledge of a DNA sequence and gene analysis can be used in several biological, medicine and agriculture research fields such as: possible disease or abnormality diagnoses, forensics, pattern matching, biotechnology, etc.

Where does DNA Sequencing, comparison and clustering come into picture?

- DNA sequence analysis can be used to identify possible errors or abnormality in a DNA sequence (e.g. in comparison with a normal one). It can be also used to predict the function of a particular gene and compare it with other “similar” genes from same or different organisms.
- If a new DNA sequence is discovered its functionality is specified based on its similarity with known DNA sequences. Such technique is used in several medical applications and research studies.

What is DNA subsequencing?

- Biologists who find a new gene sequence typically want to know what other sequences it is most similar to.
- DNA subsequencing is a way of computing how similar two sequences are.
- There are several metrics and algorithms used to decide and evaluate whether and how much two DNA sequences are similar.
- The two main algorithms are- longest common substring(LCSS) and longest common subsequence(LCS).

Sequence Alignment

- A major theme of genomics is comparing DNA sequences and trying to align the common parts of two sequences.
- If two DNA sequences have similar subsequences in common, then these sequences are homologous.
- In aligning two sequences, you consider not only characters that match identically, but also spaces or gaps in one sequence (or, conversely, insertions in the other sequence) and mismatches, both of which can correspond to mutations.
- Insertions and deletions while comparing sequences are called as indels.
- In sequence alignment, you want to find an optimal alignment that maximizes the number of matches and minimizes the number of spaces and mismatches.

Sequence Alignment

- Dedicated efforts by top Mathematicians and Computer Scientists has resulted in the development of two tools, GLobal Alignment SyStem (GLASS) and Rosetta.
- These software implementations allow for aligning pairs of homologous sequences.
- They are also used for predicting exons on a target human sequence, based on homologous comparisons from a different species.
- Both are real world applications of the Longest Common Subsequence (LCS) algorithm.

Longest Common Subsequence :

Biologists who find a new gene sequence typically want to know what other sequences it is most similar to. Finding an LCS is one way of computing how similar two sequences are: the longer the LCS is, the more similar they are.

The characters in a subsequence, unlike those in a substring, do not need to be contiguous.

Simple example -

As a simple example I've highlighted "BCBA", the longest subsequence common to both "BDCABA" and "ABCBDAB"

B D C A B A

A B C B D A B

Let's note up front that there may not be a unique LCS of two sequences (as a trivial example, "A" and "B" are both LCSes of "AB", "BA"), which is why we'll talk about **a** longest common subsequence rather than **the** longest common subsequence.

LCS Algorithm ->

```
1   m = X.length
2   n = Y.length
3   let b[1..m, 1..n] and c[0..m, 0..n] be new tables
4   for i = 1 to m
5       c[i, 0] = 0
6   for j = 0 to n
7       c[0, j] = 0
8   for i = 1 to m
9       for j = 1 to n
10          if x_i == y_j
11              c[i, j] = c[i − 1, j − 1] + 1
12              b[i, j] = "↖"
13          elseif c[i − 1, j] ≥ c[i, j − 1]
14              c[i, j] = c[i − 1, j]
15              b[i, j] = "↑"
16          else c[i, j] = c[i, j − 1]
17              b[i, j] = "←"
18  return c and b
```

The table created is shown here ->

i		j	0	1	2	3	4	5	6
		y_j		B	D	C	A	B	A
0	x_i		0	0	0	0	0	0	0
1	A		0	↑ 0	↑ 0	↑ 0	↖ 1	← 1	↖ 1
2	B		0	↖ 1	← 1	← 1	↑ 1	↖ 2	← 2
3	C		0	↑ 1	↑ 1	↖ 2	← 2	↑ 2	↑ 2
4	B		0	↖ 1	↑ 1	↑ 2	↑ 2	↖ 3	← 3
5	D		0	↑ 1	↖ 2	↑ 2	↑ 2	↑ 3	↑ 3
6	A		0	↑ 1	↑ 2	↑ 2	↖ 3	↑ 3	↖ 4
7	B		0	↖ 1	↑ 2	↑ 2	↑ 3	↖ 4	↑ 4

Algorithm to traverse the table and print the LCS is as Follows :

```
1  if i == 0 or j == 0
2        return
3  if b[i, j] == "↖"
4        PRINT-LCS(b, X, i − 1, j − 1)
5        print x_i
6  elseif b[i, j] == "↑"
7        PRINT-LCS(b, X, i − 1, j)
8  else PRINT-LCS(b, X, i, j − 1)
```

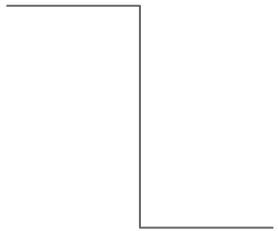
- Our project shows you basic implementations of the Needleman-Wunsch and Smith-Waterman, without optimizations, for finding global and local alignments in *O(mn)* time.
- Real-world researchers are trying to find all sequences similar to a particular sequence. If one of the similar sequences they find has a known biological function, then there is a good chance that the original sequence has a similar function because similar sequences are likely to have similar functions.
- ALIGN, FASTA, and BLAST (Basic Local Alignment Search Tool) are industrial-grade applications that find global (ALIGN) and local (FASTA and BLAST) alignments. BLAST searches large sequence databases for sequences that are similar (and possibly homologous) to a user-input sequence and ranks the results by similarity.

- BLAST first uses a process called *seeding* to find *seeds*, which are the beginnings of possible matches or hits. BLAST then uses a dynamic programming algorithm to extend the possible hits found to actual local alignments with the input sequence.

- Finally, it finds which of the matches are statistically significant and ranks them. This partly heuristic process isn't as *sensitive* (accurate) as Smith-Waterman, but it's much quicker.

Applications

- Two main families: BLAST and FASTA
- Blast is much faster than FASTA family
- LCS algorithm is base of almost all comparison and alignment algorithms
- FASTA - Smith–Waterman alg - SSEARCH - protein sequence database search
- BLAST - Needleman–Wunsch alg - NWALIGN - PYMOL image rendering



Matlab Bioinformatics Tool

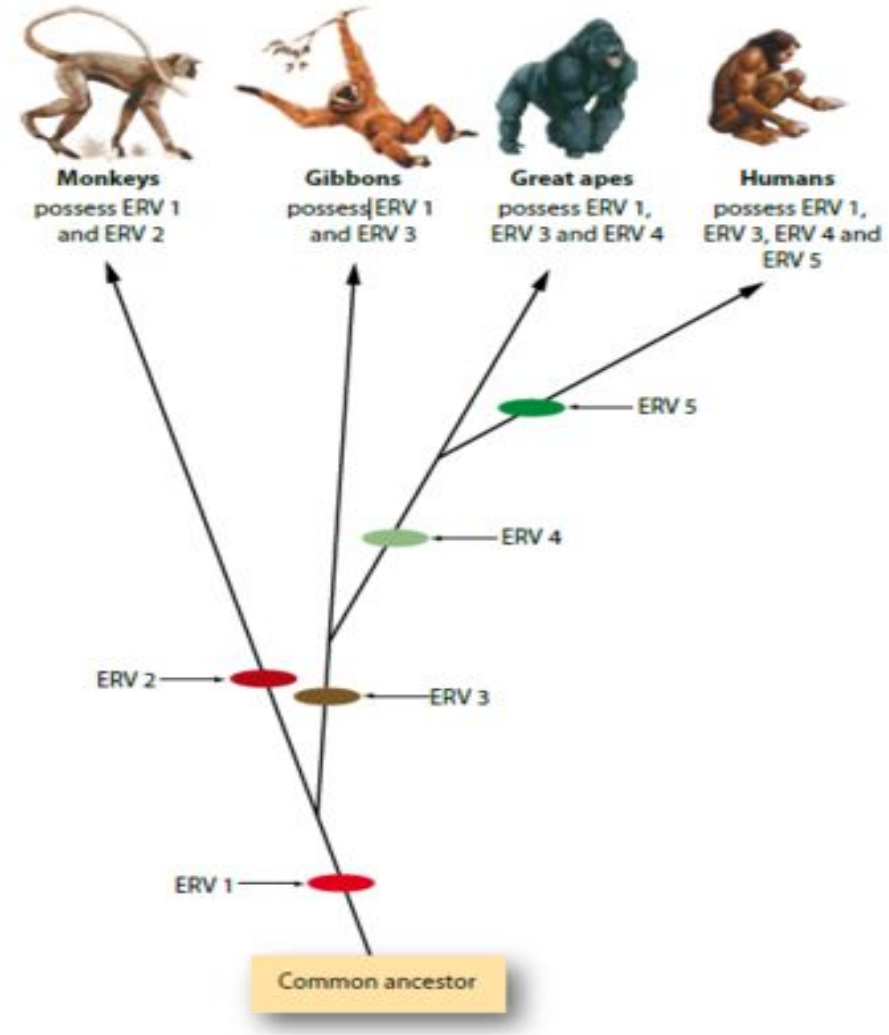
Note: here FASTA refers to FASTA tools - not FASTA file

Evolution

- `Concept:` change in genetic material over successive generation
- Creates genetically related families of organisms
- `Problem:` How to track the changes ?
- The changes in DNA are not very visible
- Homologous Sequences
- Humans and Chimpanzees have 95% - 98.5% similarity in DNA
- `Solution:` LCS can identify occurrences of similar DNA strings.
- Similar DNA strings can help identify genetically related organisms
- Similarities in DNA also provide a concrete evidence for evolution

Non functional DNA stretches such as endogenous retro viruses or ERVs may be used as evidence for evolution:

ERV: a virus which attacks reproductive chromosomes, thus leaving itsfootprint on many generations

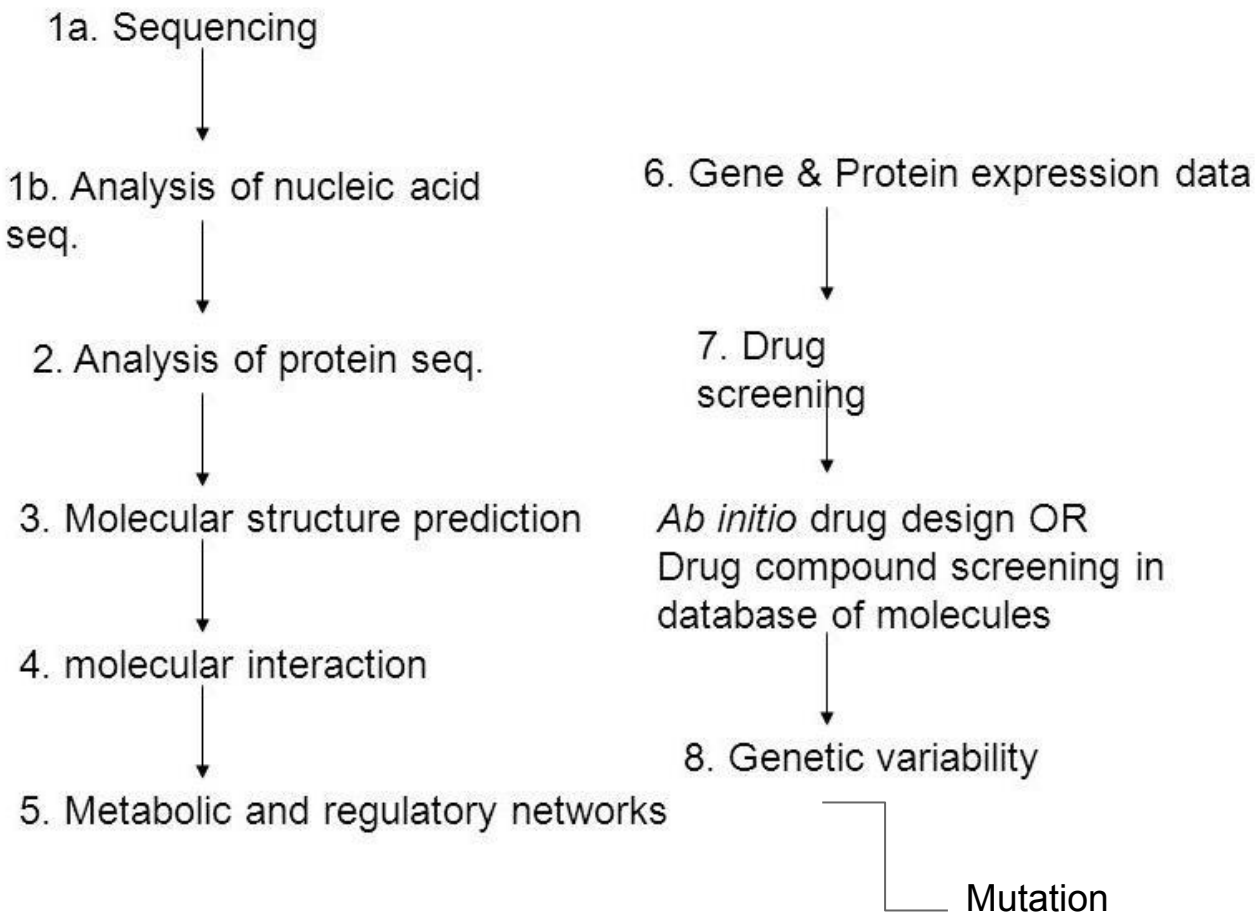


Mutation

- `Concept:` Permanent alteration in DNA sequence
- Can cause malign or benign genetic variations
- Such as change in iris color, cancer
- `Problem:` Mutations are also a cause for various diseases
- Ideal and mutated DNA may contain many identical strings
- Finding the right mutated gene is hard
- `Solution:` Find Changes in DNA *efficiently*
- Identifying mutated gene will lead to better therapy

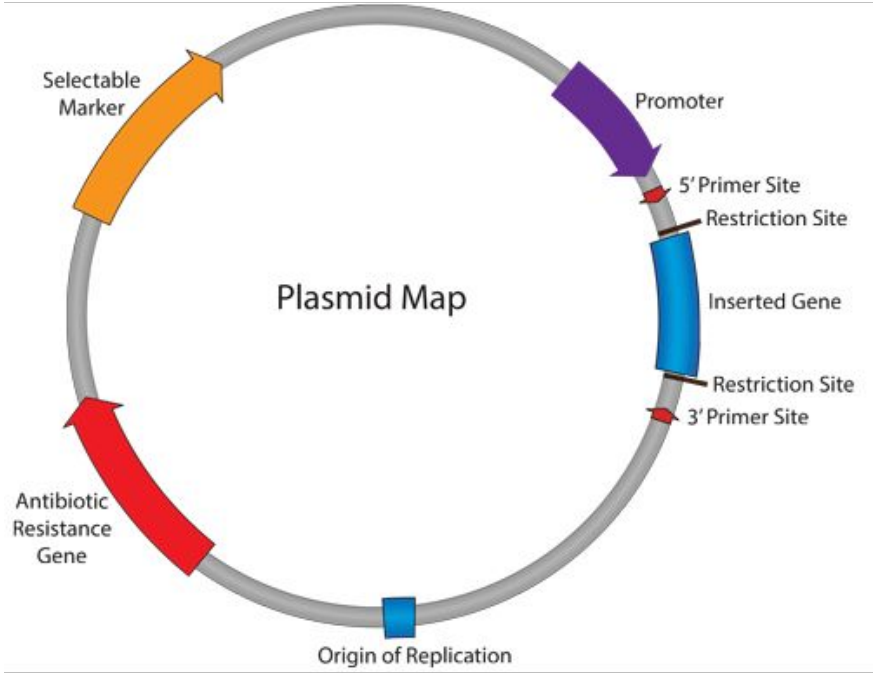
Drug Therapy Prediction by Identifying Mutations in DNA strand:

For example P53 tumor suppressor gene



Plasmid comparisons

- `Concept:` Plasmids are circular DNA strands
- Often used for storing additional Genetic info
- Upcoming DNA Data storage device
- `Problem:` Circular Data Storage
- No beginning or starting point
- `Solution:` Apply LCS to variations of plasmid
- Compare Plasmids by comparing their variations
- Comparison of Bacterial Genome



End of Presentation

Created By:

1. Shreyas Kulkarni 111608040
2. Devashish Gaikwad 111608023
3. Niramay Vaidya 111605075
4. Venkatesh Yelnoorkar 111608077
5. Atharva Jadhav 111608031

All the FASTA files can be downloaded from NCBI database