CSE115:Introduction to Biology and Chemistry for Computation

PRESENTED BY-
Tanim Ahmed
Lecturer(CSE),DIU

Lecture Outline

- Introduction
- Computer for Chemistry
- Computer for Biology
- Future scopes

Introduction

- We live in the era of computers. The word computer comes from the word “compute”, which means, “to calculate”. Despite computing it has many aspects in our life.
- The evolution of computer is still on and the ability of computers to sort, massive amount of data quickly produce useful information for almost any kind of user and makes them essential tool in modern society.

Fields of Computer

- Science(e.g. Chemical industry)
- Medicine(e.g. Bioinformatics)
- Education
- Banking
- Crime Investigation
- Entertainment
- And much more

Computers in Chemistry :

- There are several scopes of working for computer engineers in chemistry. Such as:
- **Computational chemistry**
- It is the branch of chemistry where computers are used for solving chemical problems related to simulation. It uses the methods of theoretical chemistry, incorporated into computer programs, to calculate the structures and properties of molecules, groups of molecules and solids.

Visual Models & Packages

- Many self-sufficient computational chemistry software packages exist. Some include many methods covering a wide range, while others concentrate on a very specific range or even on one method.
- For example:

- For drawing packages -*ISIS/Draw by MDL* Information Systems
- For modelling packages such as *ArgusLab*
- These software packages allow you to create your own molecular-structure

Applications of computer for chemistry in Industry

- **DCS (distributed control system)**
- A distributed control system (DCS) is **a** computerized control system for a process or plant usually with many control loops.

- **Chromatography**
- Chromatography is an analytical technique commonly used for separating a mixture of chemical substances into its individual components.

Computers in Biology :

- Computer has worked wonderfully in Medical science and biology. There are numerous of fields in biology where computer has successfully proved it’s, existence

The field of science in which **biology**, **computer science** and **information technology** merge into a single discipline

Biologists
collect molecular data:
DNA & Protein sequences,
gene expression, etc.

Bioinformaticians
Study biological questions by analyzing molecular data

Computer scientists
(+Mathematicians, Statisticians, etc.)
Develop tools, softwares, algorithms
to store and analyze the data.

Computers in Biology

- DNA sequencing
- Sequence Alignment
- Gene duplication
- DNA database searching
- Gene Therapy
- Drug development


QUESTIONS

Lecture 1.2

Role of chemistry in computer science & engineering

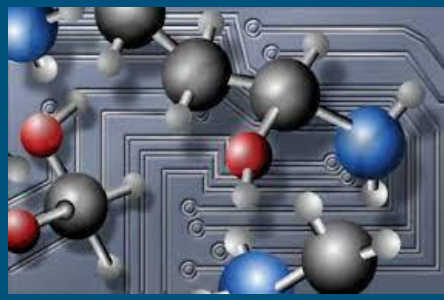
Presented By-
Faisal imran
Assistant Professor, cse, diu
Email: faisalimran.cse@diu.edu.bd

Prepared By-
Aliza Ahmed Khan,
Sr. Lecturer,,CSE, DIU
Email: aliza.cse@diu.edu.bd

CONTENTS

- Uses and effects of chemistry
- Benefits of Chemistry
- Applications of Chemistry

Uses and effects of chemistry



Computational chemistry uses result of theoretical chemistry incorporated into efficient computer programmed to calculate structure and properties of molecule.

It calculate the properties of molecule such as structure, relative energy, charge distribution, dipole moment, vibrational frequency, reactivity and other spectroscopic quantity.

Computational chemistry range from highly accurate (Ab initio method to less accurate (Semi empirical method) to very approximate (molecular mechanics).
(ab initio and semi-empirical will be discussed later)*

Cont.

In the past two decades, computational molecular modeling approaches (Leach, 2001) have emerged as important tools that **can be used to predict atomic structure, vibrational frequencies, binding energies, heats of reaction, electrical properties, and mechanical properties of organic and inorganic materials**.

Benefits of Computational Chemistry

1) It allows the medicinal chemist to use the computational power of computer for measurement of Mol. geometry , electron density , electrostatic potential, conformation analysis , different types of energies etc ...

2) Determination of structure of ligand and target through X-ray crystallography and NMR spectroscopy.

3) Docking of ligand in receptor active sites and exact measurement of geometric and energetic favorability of such interaction.

4) Comparison of various ligands through various parameters.

Applications of Computational Chemistry

https://en.wikipedia.org/wiki/Computational_chemistry

- Computational studies, used to find a starting point for a laboratory synthesis, or to assist in understanding experimental data, such as the position and source of spectroscopic peaks.
- Computational studies, used to predict the possibility of so far entirely unknown molecules or to explore reaction mechanisms not readily studied via experiments.
  Thus, computational chemistry can assist the experimental chemist or it can challenge the experimental chemist to find entirely new chemical objects.
- The prediction of the molecular structure of molecules by the use of the simulation of forces, or more accurate quantum chemical methods, to find stationary points on the energy surface as the position of the nuclear is varied.
- Computational approaches to help in the efficient synthesis of compounds.
- Computational approaches to design molecules that interact in specific ways with other molecules(e.g. Drug design and catalysis)

Watch:

https://www.youtube.com/watch?v=MA9pnR6VvBw


Thank
You

Importance of Computer in Chemical Industries

Area Covered

- Uses of Computer in chemical industries
- DCS (distributed control system)
  - Fertilizer
  - Water Treatment
  - Chemical Plant
- Chromatography

Importance of Computer in chemical industries

- Computer-Aided Chemical Engineering is being done since 1950s to the present state in which virtually all chemical engineering is computer-aided. Computer-aids are used at every stage from deciding what chemical species to make, through the conceptual design of the processes, the detailed design, the on-line control, optimization and up to the decommissioning. Computer-aids are important for assessing and minimizing environmental impacts and hazards.

- There are several areas where Computerized systems are being used in Chemical Industries. Some of them are:
- DCS (distributed control system)
    - Fertilizer
    - Water Treatment
    - Chemical Plant

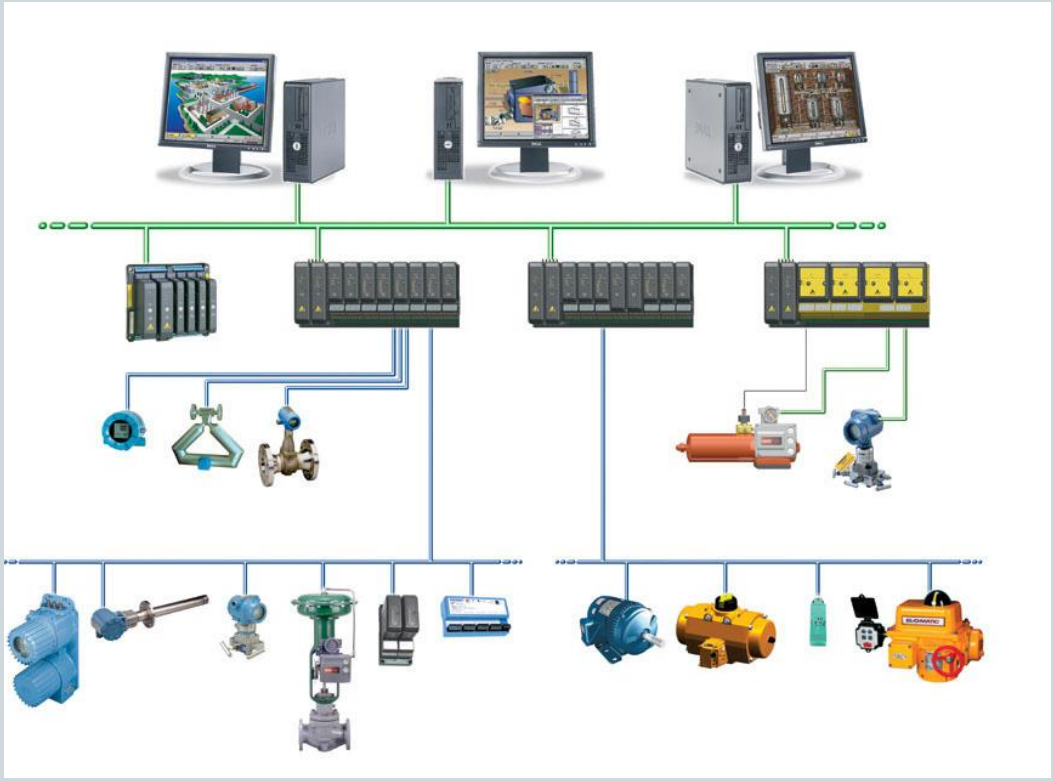
Some others areas are :

- Chemical plants
- Petrochemical (oil) and refineries
- Pulp and paper mills
- Boiler controls and power plant systems
- Nuclear power plants
- Environmental control systems
- Water management systems
- Water treatment plants
- Sewage treatment plants
- Food and food processing
- Agrochemical and fertilizer
- Metal and mines
- Automobile manufacturing
- Metallurgical process plants
- Pharmaceutical manufacturing
- Sugar refining plants

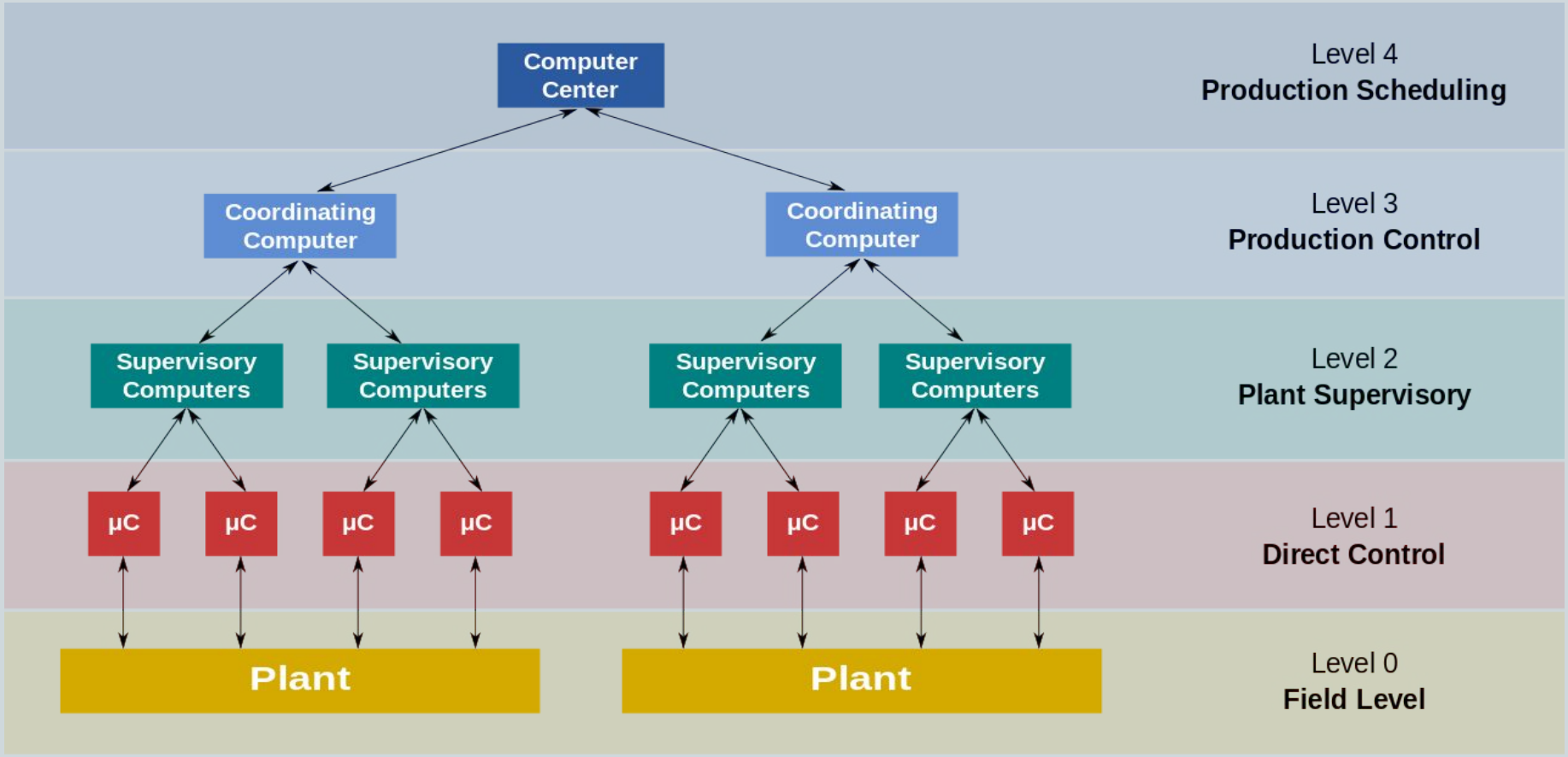
DCS (distributed control system)

- Distributed control systems (DCSs) are computer-software packages communicating with control hardware and providing a centralized human–machine interface (HMI) for controlled equipment.
- It is a central computer that autonomously coordinates the many subsystems (such as sensors and controllers) located around a plant in real-time.
- DCS are particularly important **for controlling complex processes or for large continuous manufacturing plants where top-down control and coordination is vital for efficiency**.

DCS (distributed control system)



DCS (distributed control system)

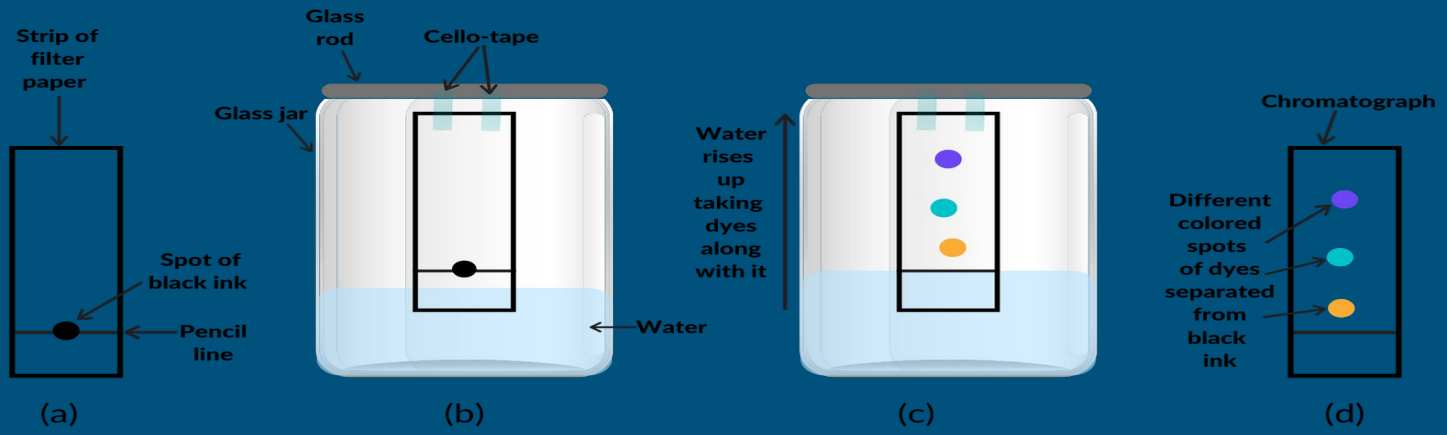


Chromatography

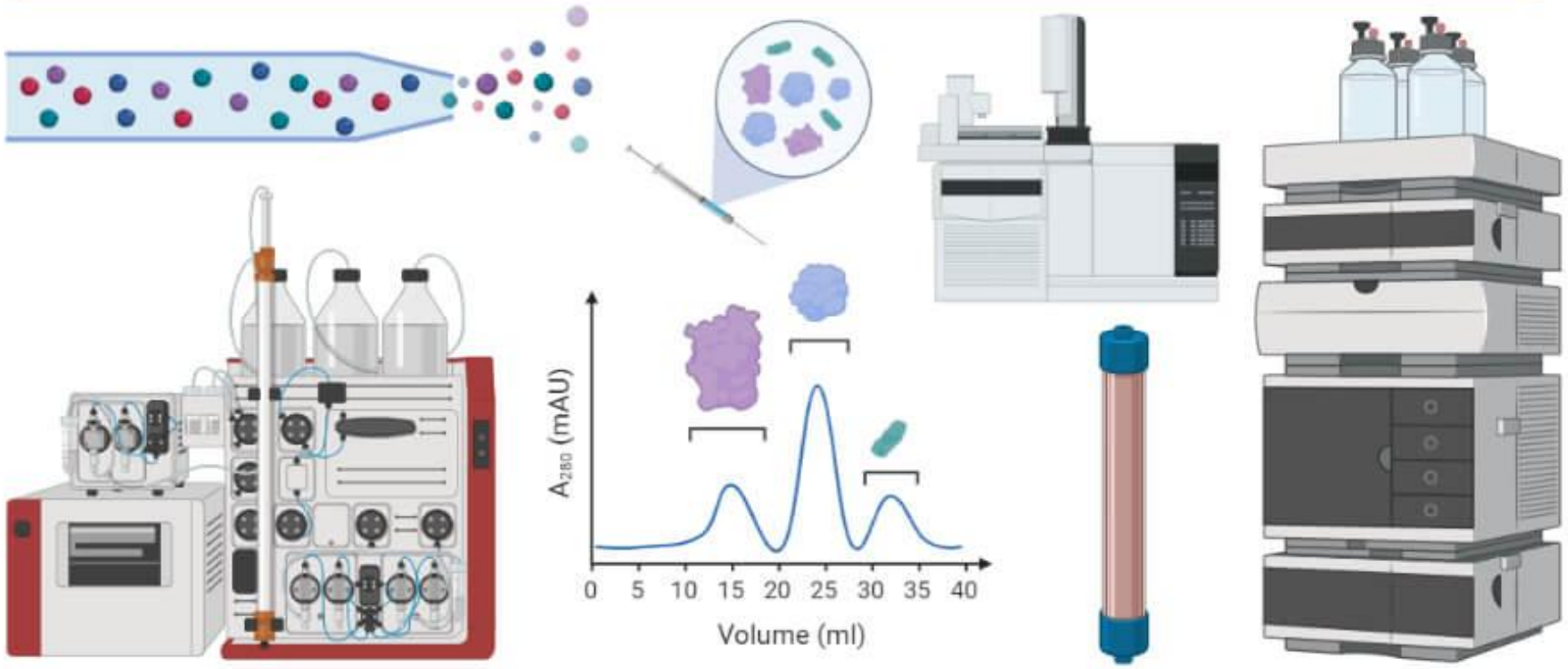
- Chromatography is **a process for separating components of a mixture**
- **Now a days computer aided systems are being used for chromatography.** Chromatography is used to separate a mixture of sample causing them to separate. Using a computer to analyze the time taken for a compound to be detected, one can know what is the compound. This can be used for detecting unknown mixture found in crime scene, mapping DNA (you can google “DNA chromatography), and so on.

teachoo

Chromatography



Types of Chromatography



Uses of Chromatography

- Let's start with some areas where it is used more often. (Not limited to)
- **Pharmaceutical industry**: In this field,(Includes Cosmetics and Herbal products too) it is mainly used to assertion purity of drugs. Identify impurities and develop chromatographic methods to quantify impurities.
- **Food and beverage industry:** In this industry, it is used to majorly identify contaminants like pesticides content in beverages or heavy metal contents in water of food stuffs.
- **Forensic Labs:** Here, chromatography is used to determine which fluids and compounds are present in human body after death or analyze blood samples to know whether he was poisoned to death etc.
- **Diagnostic Labs:** In this Labs, we determine amount of drug present in blood, urine samples etc. You would be aware of dope tests where players are tested for banned steroids.

“Tell me and I forget, teach me and I may remember, involve me and I learn.”
— Benjamin Franklin

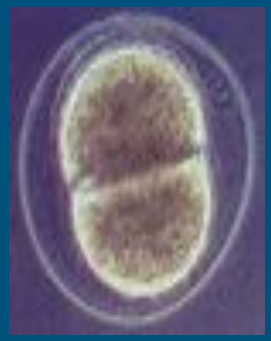
Molecular and Cellular Biology

Lecture – 2

Department of CSE, DIU

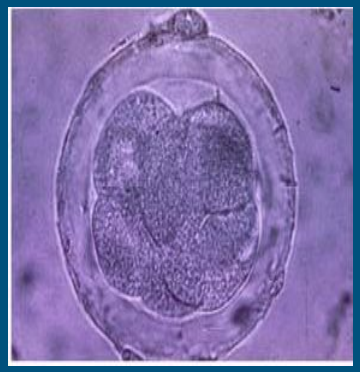
CONTENTS

What is Life made of?







1. Cell

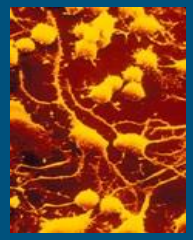
Let’s learn about Eukaryotes and Prokaryotes

Cells

- Fundamental working units of every living system
- Cell specialization in multicellular organism
- Tissues are groups of cells for a particular function
  - Fourteen major tissue types
    - Bone, muscle, nerve etc.
  - Organs are formed
- More than 200 different cell types
  - With lots of variety in every sense
  - But the genetic code is same

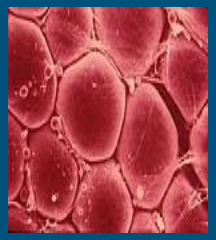
Blood

Bone



Nerve

Muscle

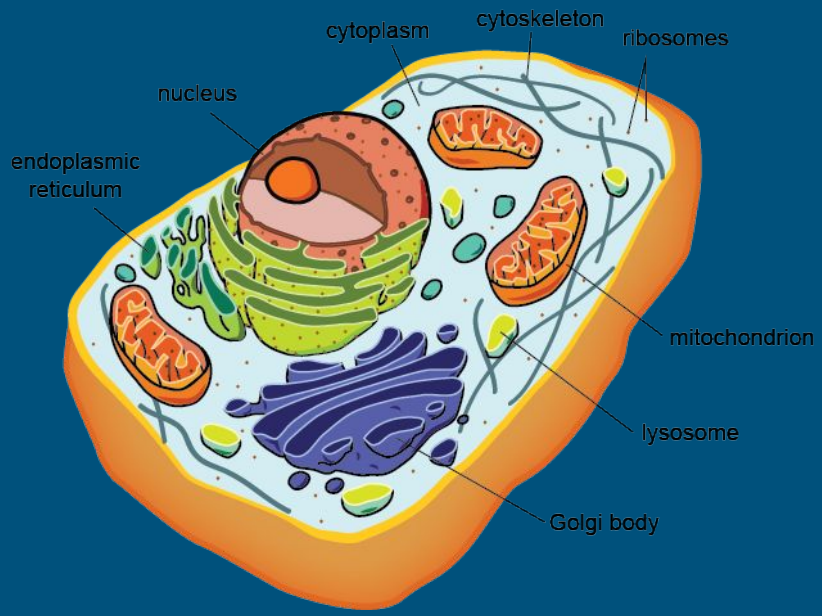


Fat

2 types of Cells

1. Eukaryotic Cells
2. Prokaryotic Cells

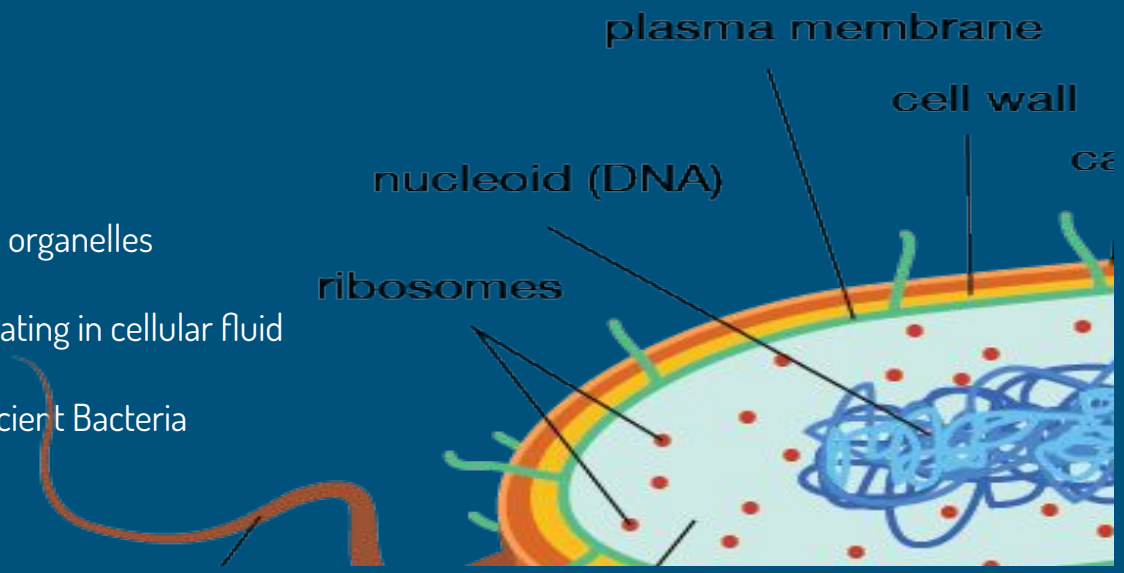
Eukaryoti c Cells



- Single or Multi Cell
- Are called Eukaryotes
- Have Nucleus
- Have membrane bounded organelles
- Have chromosomes inside Nucleus
- Seen in most of the life forms

Prokaryotic Cells

- Single Cell organism
- Are called Prokaryotes
- No Nucleus
- No other membrane bounded organelles
- One piece of rolled up DNA floating in cellular fluid
- Mostly some forms of very ancient Bacteria



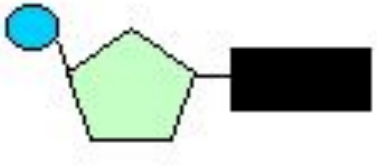
3. Nucleic Acid

All Life depends on 3 critical molecules

- DNAs
  - Hold information on how cell works
- RNAs
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein
- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body's major components (e.g. hair, skin, etc.)
  - Are life's laborers!

Building Blocks of Nucleic acids

- DNA/RNA are polymeric chain on nucleotides
- Three parts of Nucleotides
  - a nitrogenous base,
  - a five-carbon-atom sugar and
  - a phosphate group



Phosphate Molecule
Deoxyribose Sugar
Base
Adenine, Cytosine, Guanine and Thymine

Nucleic acids Bases

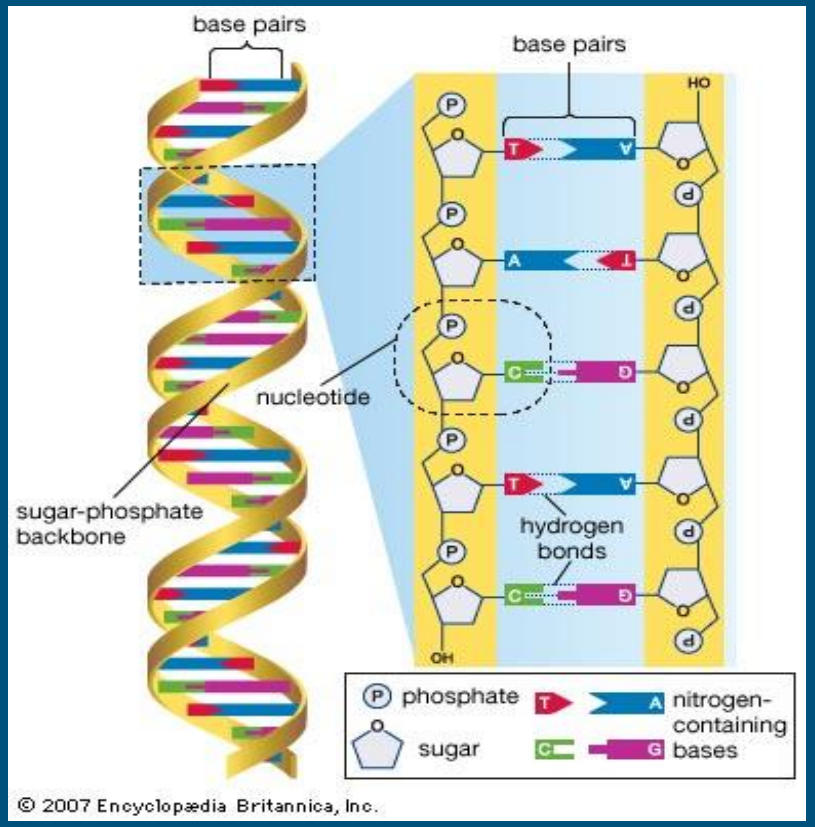
- Adenine (A),
- Guanine (G)
- Cytosine (C)
- Thymine (T)
- Uracil (U)

3.1 DeoxyriboNucleic Acid (DNA)

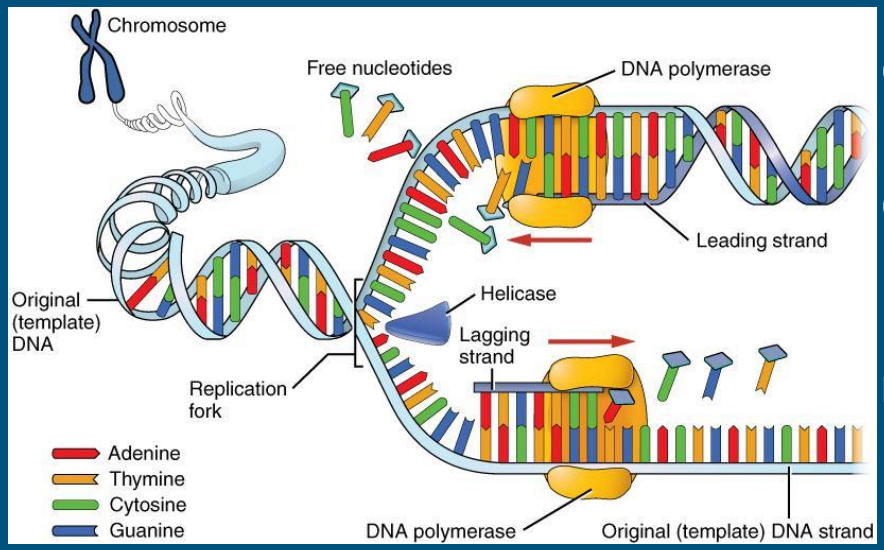
Carrier of genetic instructions

DNA Structure

- **Double Helix Structure (Watson and Crick, Nature 1953)**
- Two complementary antiparallel strands, one runs from 5' to 3' end and another runs from 3' to 5' end
- 3 major parts – Nitrogenous Base, 5-Carbon Deoxyribose Sugar and Phosphate Group
- Four nitrogenous bases – Adenine (A), Cytosine (C), Guanine (G), Thymine (T)
- A-T is Double Hydrogen Bond and G-C is Triple Hydrogen Bond
- DNA is more stable than RNA due to its Deoxyribose Sugar Structure



DNA Replication



▹Initiation
- Helicase enzyme unwinds DNA strands
- Replication fork is created
- RNA Primer is created by Primase enzyme
- Primer is starting point of elongation

▹Elongation
- New DNA Strand grows 1 base at a time as complimentary of leading strand (5' to 3')
- DNA Polymerase enzyme controls it
- Complimentary strand of lagging strand is created in small fragments called Okazaki Fragments (3' to 5')

▹Termination
- Exonuclease enzyme removes all the primer sequences from new strands
- Again, DNA Polymerase fills the gaps
- DNA Ligase enzyme seals all the gaps

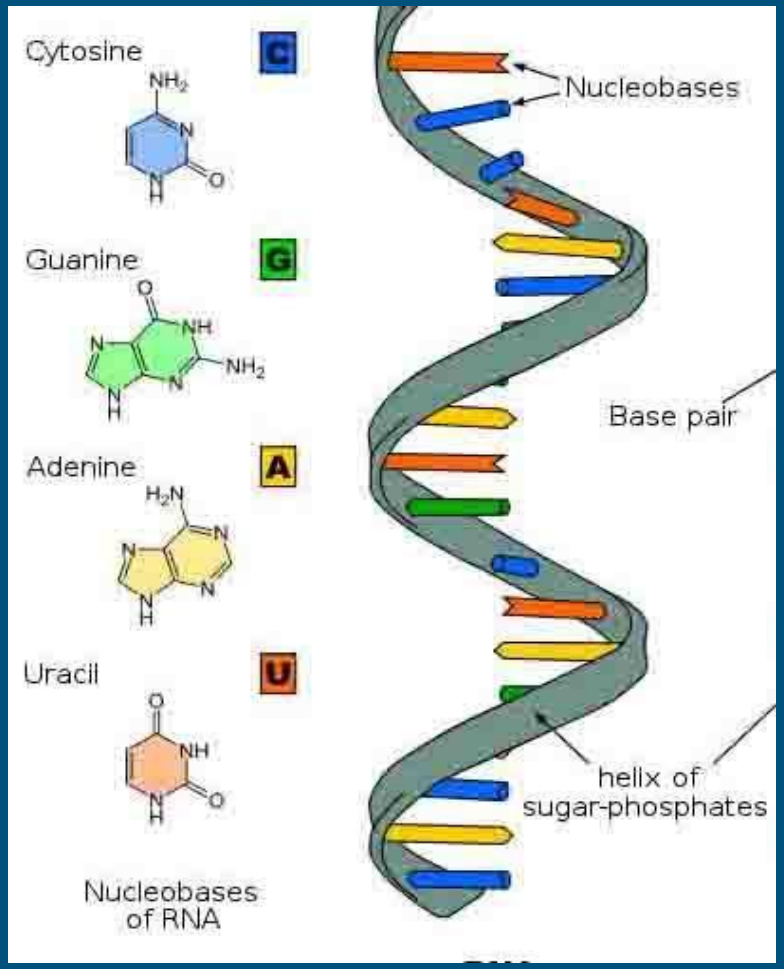
*** DNA Replication is Semi-Conservative, because, in new sets of DNA, one strand is newly created but the other strand comes from the ancestor.**

3.2 RiboNucleic Acid (RNA)

Protein Coding and Carrier

RNA Structure

- Single Helix Structure
- Single Strand which generally runs from 5' to 3'
- 3 major parts – Nitrogenous Base, 5-Carbon Ribose Sugar and Phosphate Group
- Four nitrogenous bases – Adenine (A), Cytosine (C), Guanine (G), Uracil (U)
- A-U is Double Hydrogen Bond and G-C is Triple Hydrogen Bond
- RNA is less stable than DNA due to its Ribose Sugar's structure



RNA Types

Messenger RNA (mRNA)
Carries a genes coding message for protein from Nucleus to Ribosome

Transfer RNA (tRNA)
Transfers specific amino acid sequence to ribosome to form Protein

Ribosomal RNA (rRNA)
Protein and rRNA combinedly forms ribosome

Non-Coding RNA
Not translated into protein. Ex – tRNA, rRNA

Catalytic RNA
Catalyze chemical reaction.

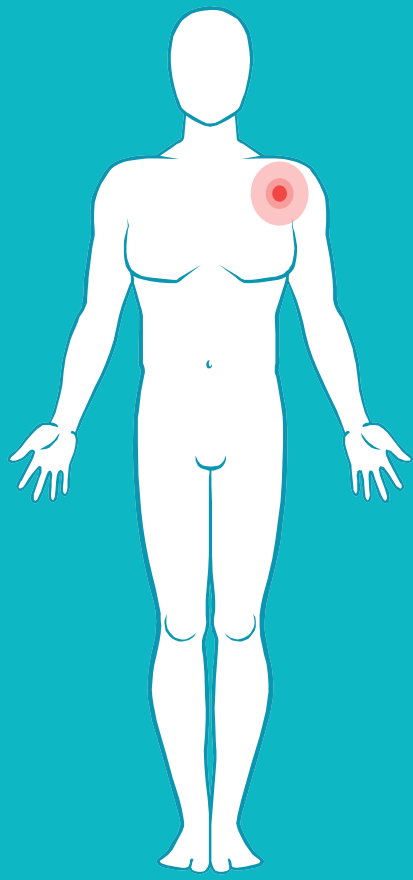
Double Stranded RNA
Contains complementary strands like DNA. Induces gene expression.

Reference Video

https://youtu.be/C1CRrtkWwu0
https://youtu.be/TNKWgcFPHqw


Thank
You

DNA Sequencing



Lecture – 4

Taslima Ferdaus Shuva, Sr. Lecturer, Department of CSE, DIU

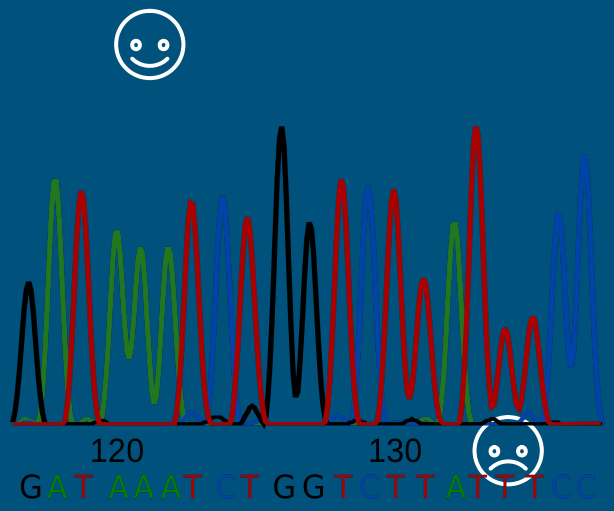
CONTENTS

1. DNA Sequencing

Determining nucleotide sequences

DNA Sequencing

- DNA sequencing is the process of determining the precise order of nucleotides (A, T, G, C) within a DNA molecule.

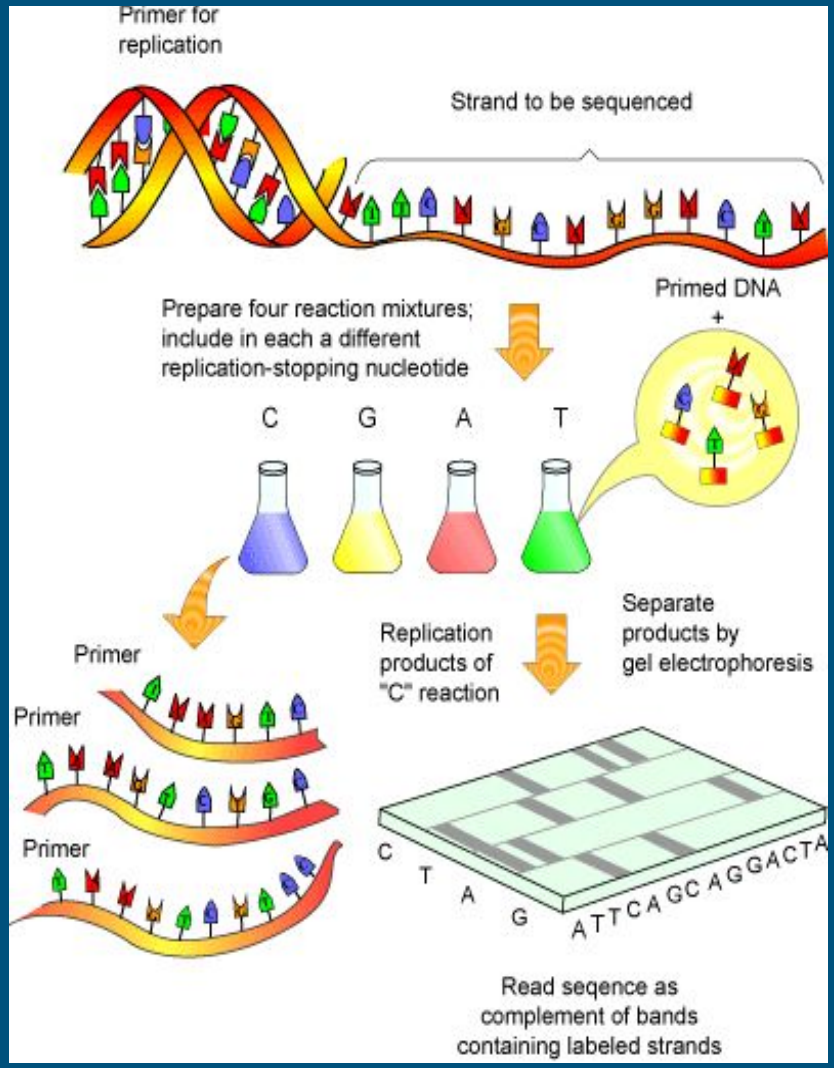


2. First Generation Sequencing

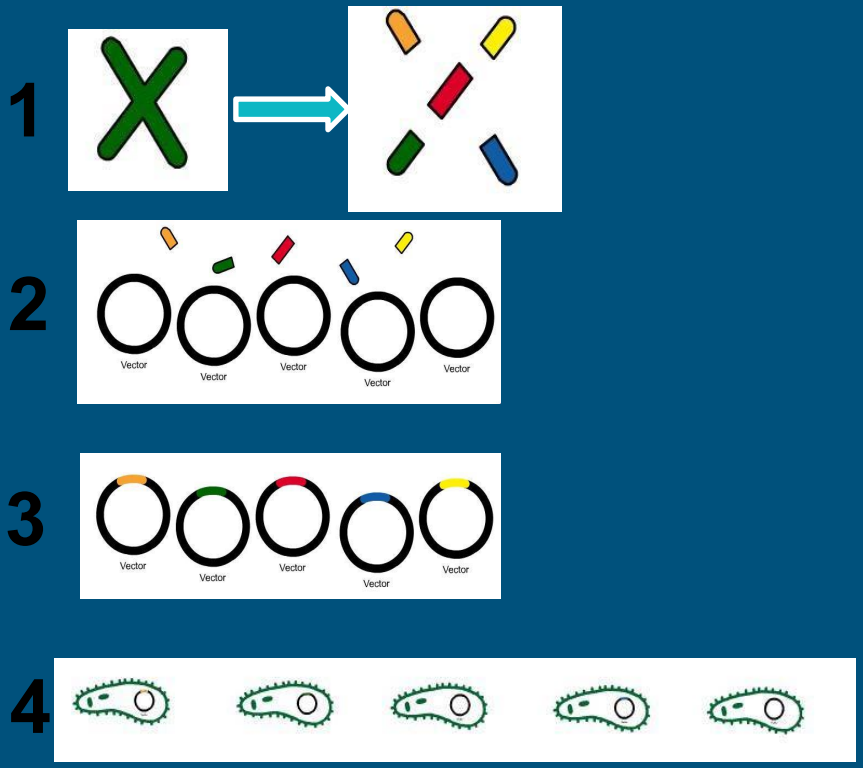
Predominant method for sequencing for decades

Sanger Method

- Developed by Frederick Sanger in 1977
- Most popular and predominant method for DNA Sequencing for decades
- Can read up to 2000 bps
- Slow and expensive
- Labor intensive
- Human Genome Project was completed using Sanger Sequencing



Step 1 - DNA Preperation

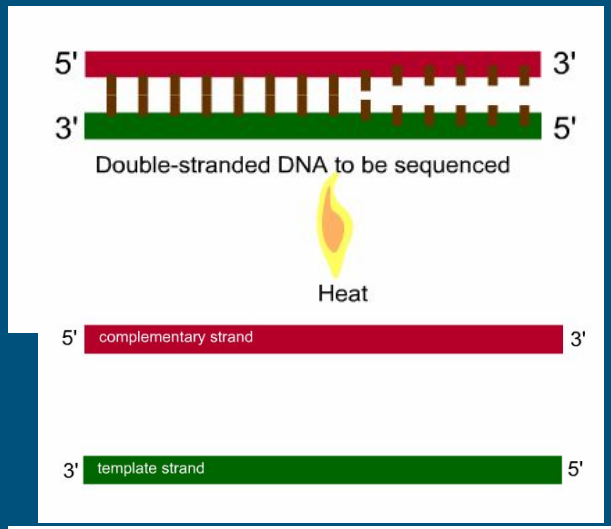


- Cut DNA into a smaller piece for sequencing
- Insert into Plasmid
- Insert Plasmid inside Bacteria Cell and let it multiply
- Extract all the necessary Plasmids and from Plasmid, isolate the DNA for sequencing

Step 2 –Sequencing Reaction

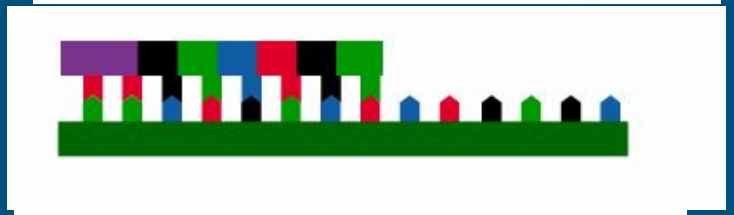
- Strand Separation
  - Heat DNA in 96° C (denaturation)
- Primer Annealing
  - Lower temperature to 50° C (annealing)
  - Primer binds to DNA
- Primer Extension
  - Increase temperature to 60° C
  - DNA Polymerase binds to Primer
  - Add complementary bases (dNTP) after Primer until terminator base is added (ddNTP)
- Termination
  - Terminate chain after ddNTP is added
  - ddNTP is fluorescently labelled (different colors for A, T, G, C)

1.



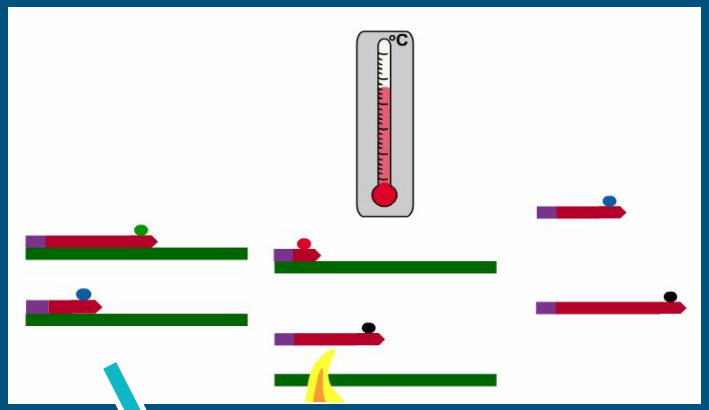
2.

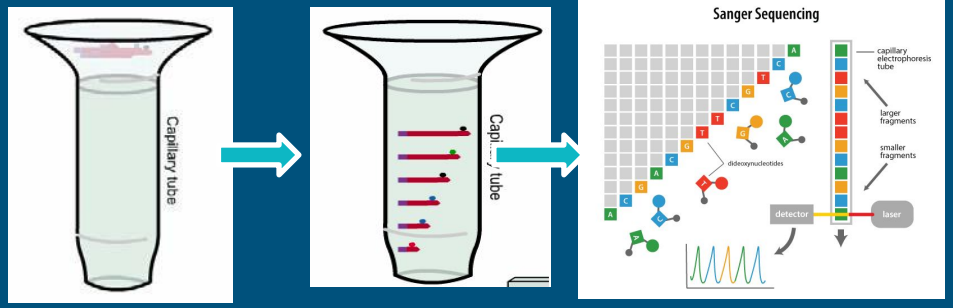
3.



4.

Step 3 – Electrophoresis in Capillary





- Sort the newly synthesized DNA strands by length (নতুন সংশ্লেষিত ডিএনএ স্ট্র্যান্ডগুলি দৈর্ঘ্য অনুসারে সাজান)
  Strands are loaded inside a capillary tube
- An electrical negative charge pulls positively charged DNA strands through the capillary
- Emerged strands pass through a laser beam that excites the ddNTP fluorescent dye at the end of each strand
- Beam causes dye to glow in a specific wavelength/color which is captured by photocell and stored in a computer
- Computer than maps each color to each nucleotide sequentially and generates final sequence output

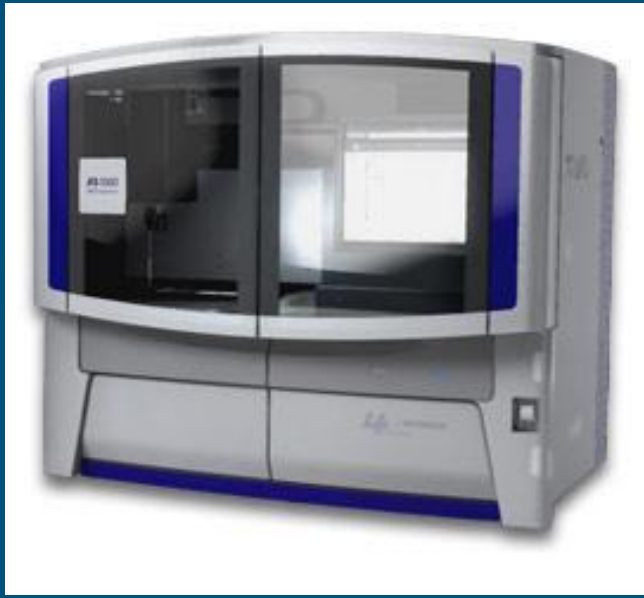
3. Second / Next Gen Sequencing

Less Costly methods, mostly Short Read Sequences, High number of reads

454/Roche (2005)

- Pyrosequencing technique
- Long Read Sequencing (length up to 700 bps)
- Accuracy 99.9%
- Can sequence up to 1 Million reads/run
- Fast (around 24 hours/run)
- Expensive (costs around $10 per 1 million base)

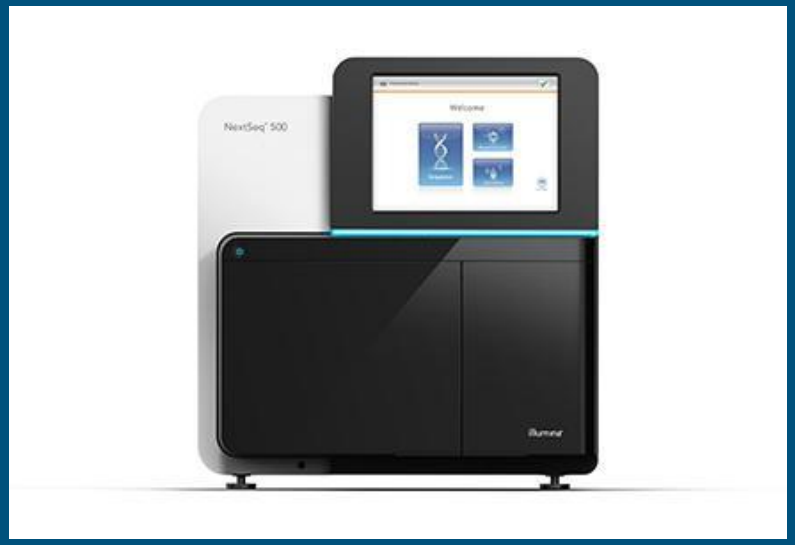
ABI SOLiD (2006)



- SOLiD (Sequence by Ligation)
- Short Read Sequencing (length up to 100 bps)
- Accuracy 99.9%
- Can sequence up to 1.4 Billion reads/run
- Time around 1-2 weeks, Slower than other sequencers
- Cheap (costs around $0.13 per 1 million base)

Illumina / Solexa (2007)

- Sequencing by Synthesis
- Short Read Sequencing (length up to 300 bps)
- Accuracy 99.9%
- Can sequence up to 3 Billion reads/run
- Moderately Slow (around 1-11 days/run)
- Expensive Equipment, run cost is low (costs around $0.05-$0.15)



4. Third / Next-Next Gen Sequencing

Long reads, Higher error rate

Pacific Biosciences (PacBio)

- Single Molecule Real Time Sequencing
- Long Read Sequencing (length up to 40,000 bps)
- Accuracy 87%
- Can sequence up to 500-1000 Mega reads/run
- Time around 30 minutes – 4 hours, Faster
- Expensive Equipment, run cost is low (costs around $0.13-$0.60)

Oxford Nanopore

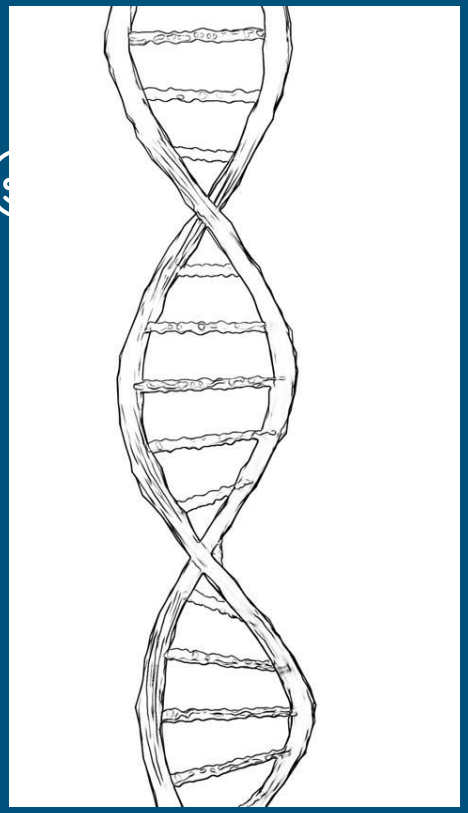
- Nanopore sequencing
- Very Long Read Sequencing (length up 500 kb), Portable
- Accuracy 92-97%
- Depends on read length selected by user
- Time around 1 minutes – 48 hours, Faster
- Expensive Equipment, run cost is low (costs around $500-$999 per flow cell)

5. Miscellaneous Terms

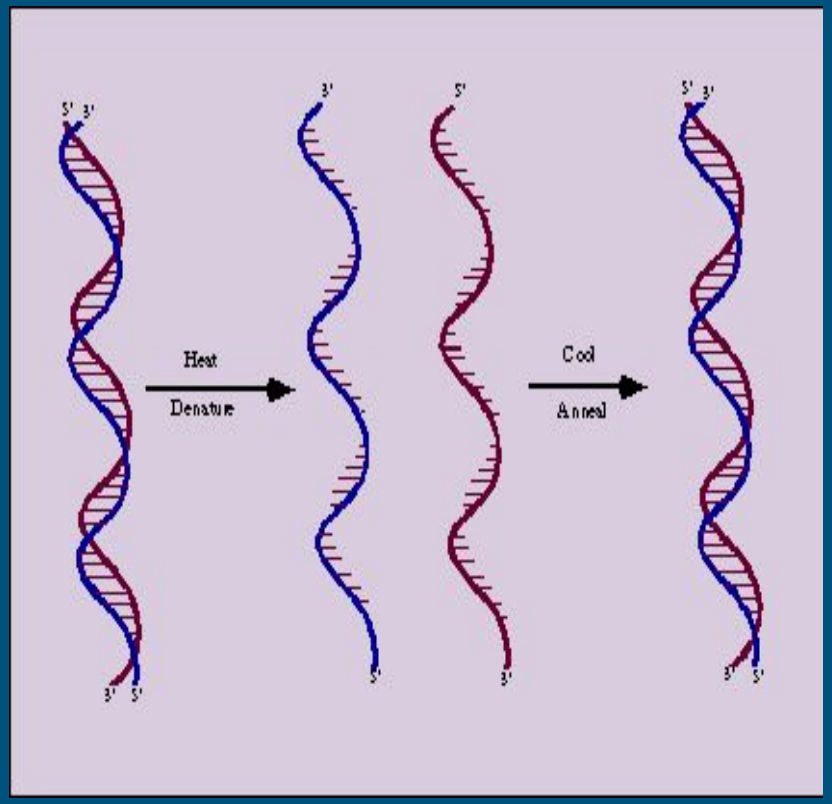
Some comparisons, terms etc.

Oligonucleotide

- Short sequences of DNA or RNA
- Typically less than 20bp
- Oligonucleotide of ‘k’ bases length is called k-mer.



Denaturation and Annealing



Denaturation

- Energy of heat pull apart two DNA strands
- Happens at a critical temperature denoted T_m

Annealing

- Decrease temperature, and strands are joined back together
- Only complementary bases will bond

Plasmid

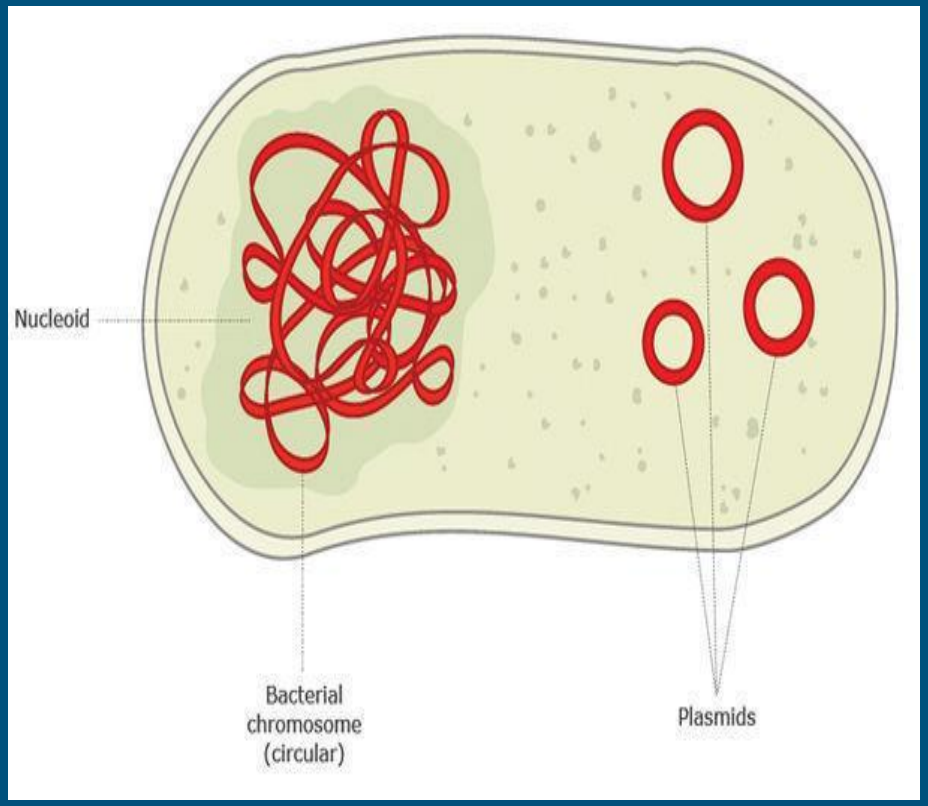
- Small, circular piece of DNA often found in bacteria.
- Sizes of 2.5-20 kb
- Plasmid using method -

    * Isolate them in large quantities

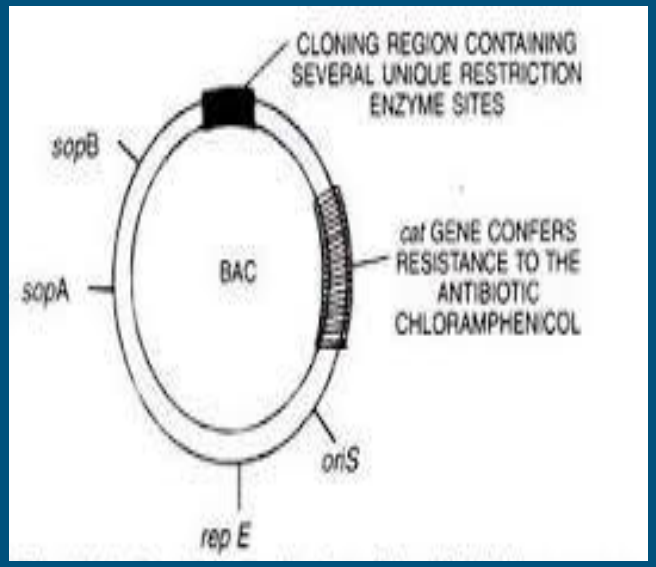
    * Cut and splice them, adding whatever DNA needed

    * Put them back into bacteria, where they'll replicate along with the bacteria's own DNA

    * Isolate them again - getting billions of copies of whatever DNA was inserted into the plasmid



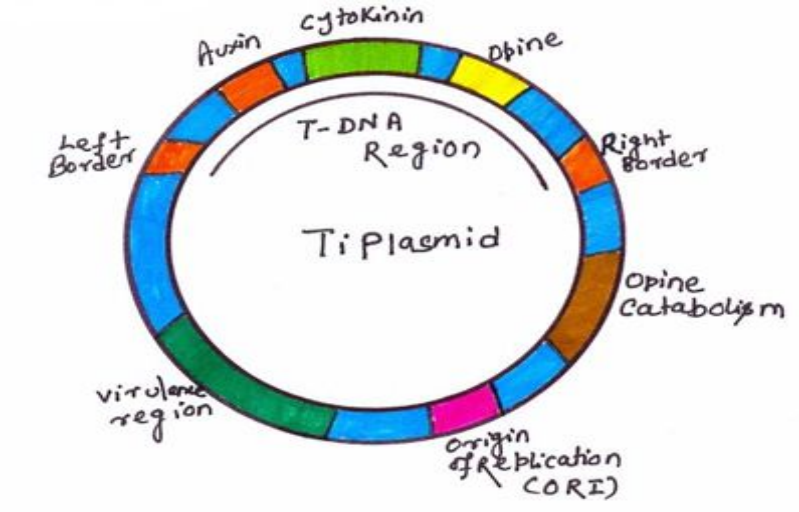
Bacterial Artificial Chromosome (BAC)



- **Used like a plasmid**
- **BACs carry DNA from humans or mice or any other living being, and is inserted into a host bacterium for replication**
- **BAC is artificially constructed, unlike Plasmid**

Cloning Vector

- A cloning vector is a small piece of DNA, taken from a virus, a plasmid, or the cell of a higher organism, that can be stably maintained in an organism, and into which a foreign DNA fragment can be inserted for cloning purposes.



8%

Of Human DNA is made of Ancient Viruses

700 Terabytes

Data can be stored in 1gm DNA

50 Years Time

Type entire human genome at a speed of 60

99.9%

Human DNA is identical, 0.01% creates human diversity

TO BE CONTINUED

Impressed?

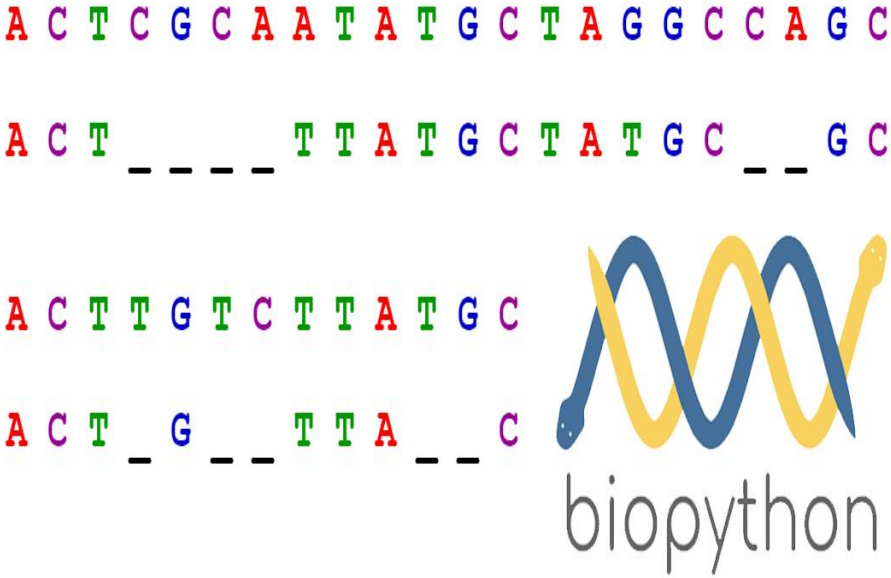
Youtube Links

- Sanger Sequencing – https://www.youtube.com/watch?v=0NGdehkB8jU

Sequence Alignment

Lecture – 5

Department of CSE, DIU



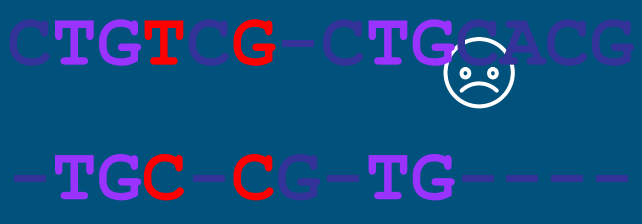
CONTENTS

1. Sequence Alignment

Why and how align sequences

Sequence Alignment

A way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences



Why align sequences?

- Useful for discovering
  - Functional
  - Structural and
  - Evolutionary relationship
  - For example
    - To find whether two (or more) genes or proteins are evolutionarily related to each other
    - Two proteins with similar sequences will probably be structurally or functionally similar

2. Sequence Alignment Methods

Pairwise and Multiple

Pairwise Sequence Alignment

- A pair of sequences as input
- Align them in such a way that, for that particular alignment the assumed region of similarity produces higher score than all the other alignments
- Methods
  - Global Alignment (Needleman-Wunsch)
  - Local Alignment (Smith-Waterman)

```
CTGTCGCTGCACG--
-------TGC-CGTG
```

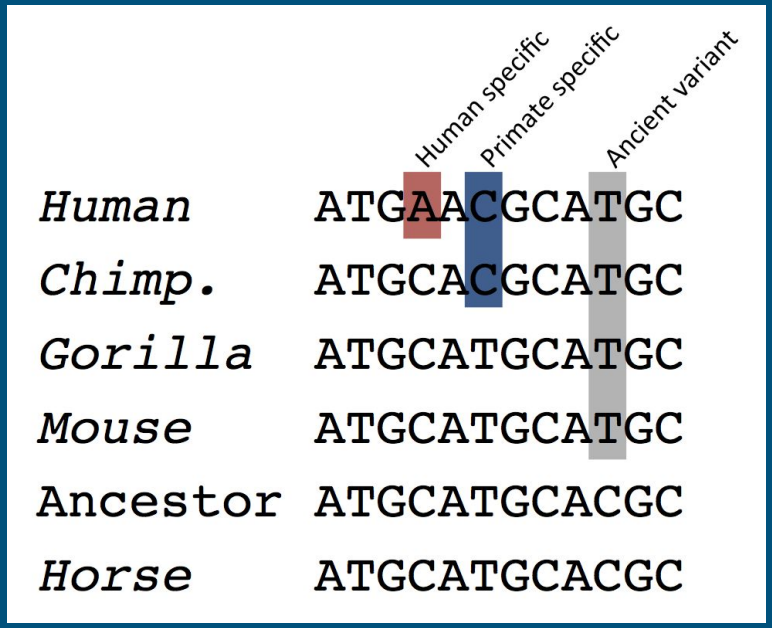
Pairwise Sequence Alignment

- Idea:

  Display one sequence above another with spaces inserted in both to reveal similarity

```
A: C A T - T C A - C
   |   |     | |   |
B: C - T C G C A G C
```

Multiple Sequence Alignment



- Three or more than three sequences as input
- Align all the sequences altogether in such a manner that the alignment produces highest score

3. Pairwise Sequence Alignment

Global and Local methods

Global Alignment (Needleman-Wunsch)

3 Major Steps
- -Create 2D Matrix
- -Trace back
- -Final Alignment

Create 2D Matrix
- Row x Col 2D matrix draw (Row , Col size of seq1 and seq2 respectively)
- Place 2 seqs as Row and Column Header
- Cell (0,0) = 0
- Cell (0,1) to Cell (0,Column) and Cell (1,0) to Cell (Row,0) value = delete gap value from previous cell value
- For other cell values, follow equation in (1)

Trace back
- Start from Cell (Row, Col)
- Go back up to Cell (0,0)

Final Alignment
- Start from Cell (Row, Col)
- If then place character in both seq
- If then character in start seq & gap in end seq

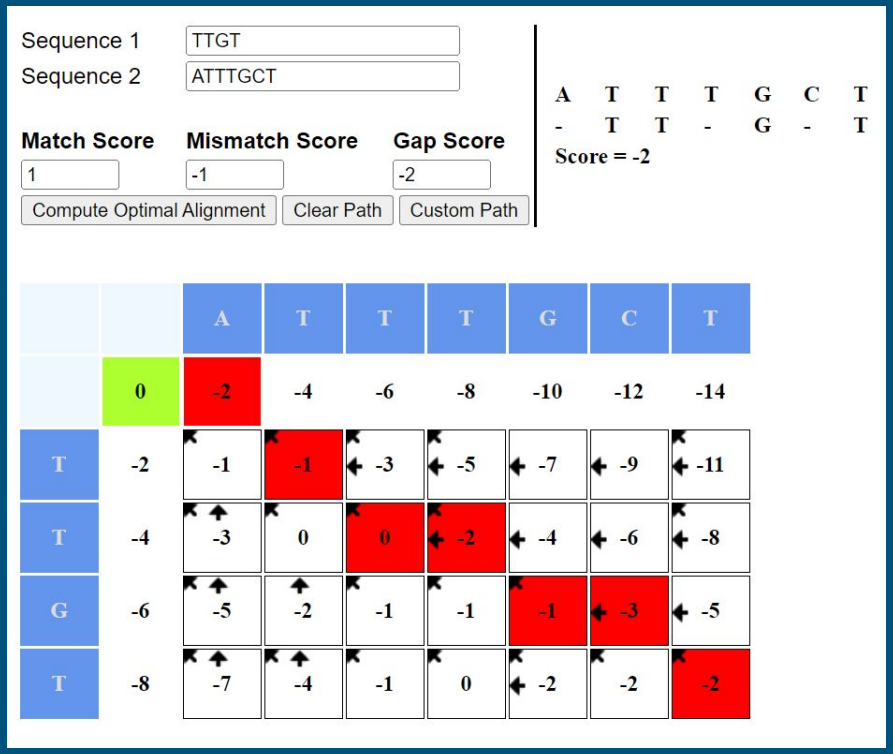
Global Alignment (Needleman-Wunsch) - Example

Input
- seq1 = TTGT
- seq2 = ATTTGCT

Scoring Scheme
δ(x, x) = 1 (Match)
δ(x,-) = -2 (Gap)
δ(x, y) = -1 (Mis match)

$$V_{i,j} = \max \begin{cases} V_{i-1,j} + \delta(s_i, -) \\ V_{i,j-1} + \delta(-, t_j) \\ V_{i-1,j-1} + \delta(s_i, t_j) \end{cases}$$

Eq. 1: Cell Value



Local Alignment (Smith-Waterman)

3 Major Steps

-Create 2D Matrix

-Trace back

-Final Alignment

Create 2D Matrix

- Row x Col 2D matrix draw (Row , Col size of seq1 and seq2 respectively)
- Place 2 seqs as Row and Column Header
- First Row, First Column all value = 0
- For other cell values, follow equation in (2)

Trace back

- Start from each Cell which has the maximum value in the entire matrix
- Go back up to the Cell where first time 0 occurs

Final Alignment

- Start from each Cell with max value
- If then place character in both seq
- If or then character in start seq & gap in end seq

Local Alignment (Smith-Waterman) - Example

Input
- seq1 = TCGT
- seq2 = GATTCGT

Scoring Scheme
δ(x, x) = 2 (Match)
δ(x,-) = -3 (Gap)
δ(x, y) = -2 (Mis match)

$$A[i,j] = \max \begin{cases} A[i,j-1] + \text{gap} \\ A[i-1,j] + \text{gap} \\ A[i-1,j-1] + \text{match}(i,j) \\ 0 \end{cases}$$

Eq. 2: Cell Value

Sequence a: TCGT

Sequence b: GATTCGT

Scoring in s: Match 2 Mismatch -2 Gap -3

Hint:
For similarity maximization,
match scores should be positive and all other scores lower.

Recursion:

$$S_{i,j} = \max \begin{cases} S_{i-1,j-1} + s(a_i, b_j) \\ S_{i-1,j} + s(a_i, -) \\ S_{i,j-1} + s(-, b_j) \\ 0 \end{cases} = \max \begin{cases} S_{i-1,j-1} + 2 & a_i = b_j \\ S_{i-1,j-1} + -2 & a_i \neq b_j \\ S_{i-1,j} + -3 & b_j = - \\ S_{i,j-1} + -3 & a_i = - \\ 0 \end{cases}$$

Output:

S		G_1	A_2	T_3	T_4	C_5	G_6	T_7
	0	0	0	0	0	0	0	0
T_1	0	0	0	2	2	0	0	2

Results
You can select a result to get the related traceback.

Methods of Computational Chemistry

Prepared By-
Faisal Imran
Assistant Professor
CSE, DIU

Molecular modeling method

The three dimensional shape of both ligand and target site may be determined by X-ray crystallography or computational method.

The most common computational methods are based on either molecular or quantum mechanics.

Both this approaches produce equation for total energy of the structure.

Computational method

- There are two main types method depending on the starting point theory.

- Classical method :-

  Are those method use Newton mechanics to model molecular system.

- Quantum chemistry method:-

  Which makes use of Quantum mechanics to model the molecular system. This method used different type of approximation to solve Schrödinger's Equation.

- **Classical Methods**

  1. Molecular Mechanics
  2. Molecular Dynamics.

- **Quantum Mechanics Methods**

  1. Semi empirical Methods.
  2. *Ab initio* Methods.
  3. Density functional Theory.

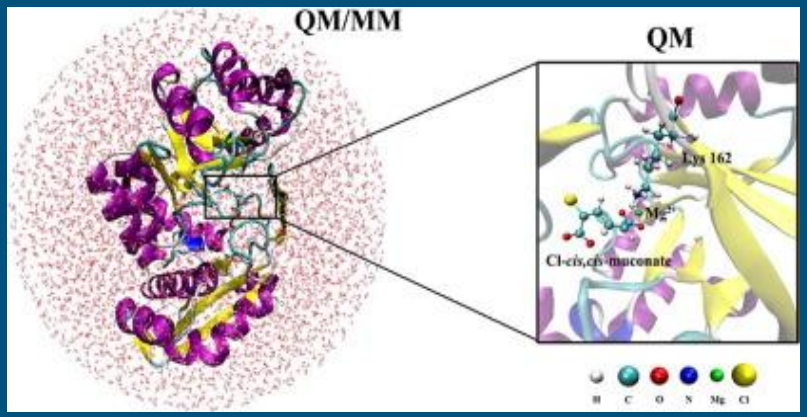
What is Computational Chemistry?

Application of computational methods and algorithms in chemistry

- Quantum Mechanical
  i.e., via *Schrödinger* Equation
  also called *Quantum Chemistry*
- Molecular Mechanical
  i.e., via *Newton*'s law F=ma
  also *Molecular Dynamics*
- Empirical/Statistical
  e.g., *QSAR*, etc., widely used in clinical and medicinal chemistry

$$-i\hbar\frac{\partial}{\partial t}\Psi = \hat{H}\Psi$$

Focus Today



Differences Between Molecular Mechanics & Quantum Mechanics

Molecular Mechanics Vs Quantum Mechanics

A. Quantum mechanical methods

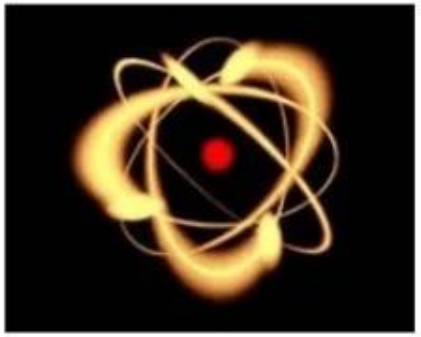
- **deal with electrons in a system, and**
- **the calculations are very MUCH time consuming.**

B. Molecular mechanics (force-field methods) on the other hand

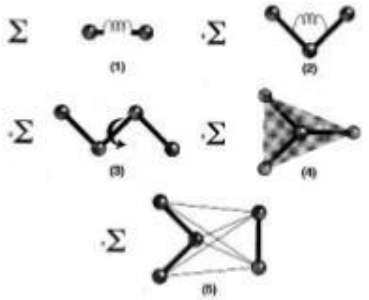
- **ignore the electronic motions and**
- **calculates the energy of the system as a function of the nuclear position only.**
- **the calculations are very LESS time consuming.**

However, molecular mechanics cannot provide answers that rely on the electronic distribution of a molecule.

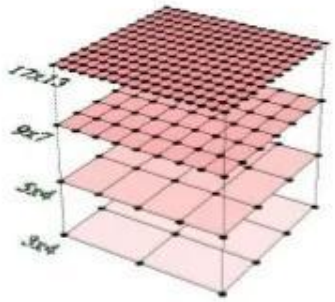
Computational Chemistry



Quantum Mechanics (QM)



Molecular Mechanics (MM)



Hybrid QM / MM

Semi-empirical (SE)

Molecular Mechanics

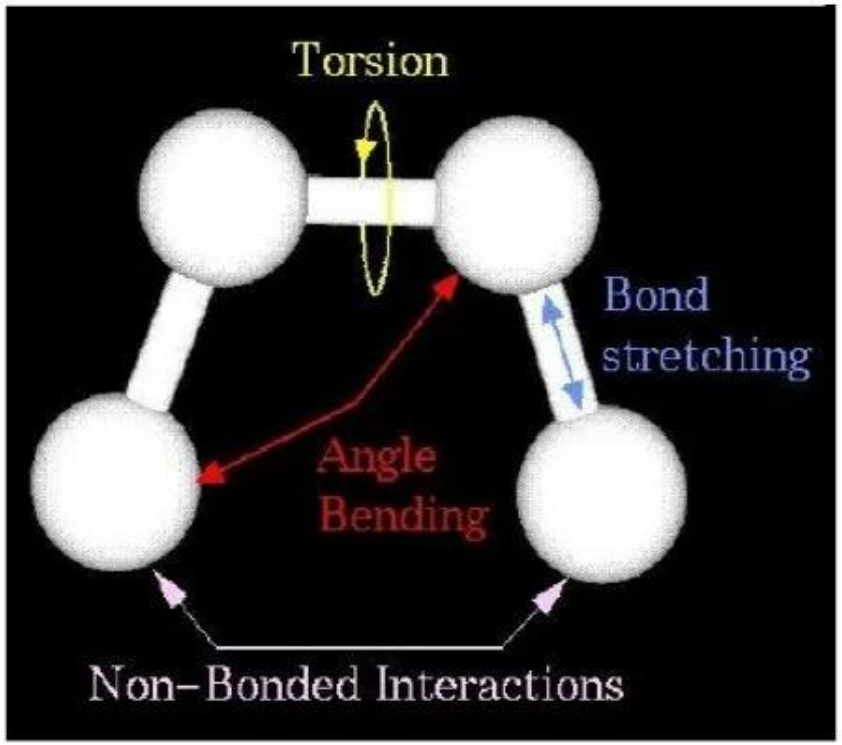
- Molecular mechanics programs use equations based on classical physics to calculate force fields.

- Atoms treated as spheres, bonds as springs and electron are ignored.

- It assume that the total potential energy (E_{total}) of molecule is given by sum of all the energies of attractive and repulsive forces between atom in structure.

The molecular mechanics equation

$$E = E_B + E_A + E_D + E_{NB}$$



E_B = The energy involved in the deformation bond either by stretching or compression.

E_A = The energy involved in the angle bending .

E_D = The torsional angle energy.

E_{NB} = The energy involved in the interaction between atoms that are not directly bonded.

Force Field

- Force field refers to calculation of the interaction and energies between different atoms between bond stretching, angle bending, torsional angle and non-bonded interaction.
- Force field ignores the electronic distribution while Quantum mechanics considers electronic distribution of molecule.

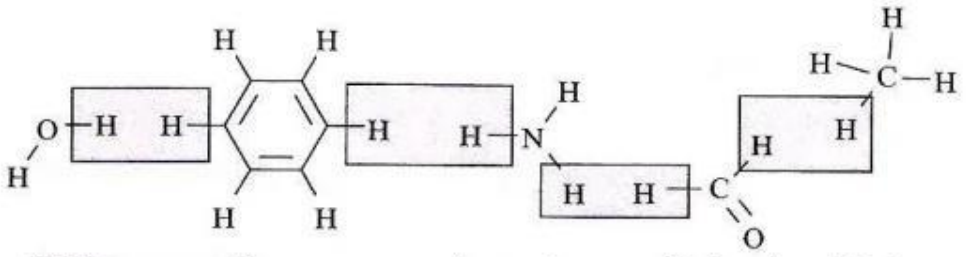
Classical empirical force field

1) AMBER(Assisted Model Building and energy Refinement)
2) CHRAM (Chemistry at Harvard Macromoleculer Mechanics)
3) CVFF(Consistent Valence Force Field)

Molecular Model Using Molecular Mechanics

- The molecular models are created by either using an existing commercial force field computer program or assembling a model from structural fragments held in the database of molecular modeling program.

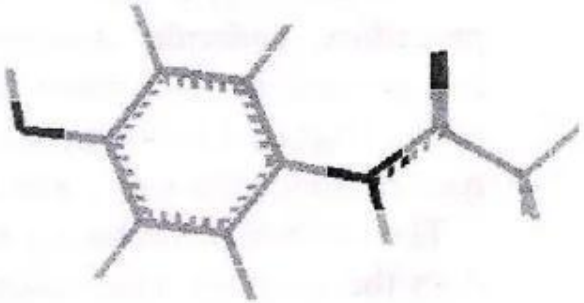
Step 1: the selection of the structure fragments from the database of the INSIGHT II program. The molecule with the relevant functional group and/or structure is selected.

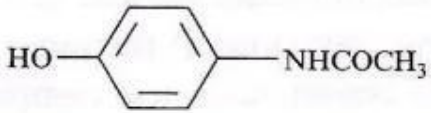


The INSIGHT II models of these structures.

Step 2: The fragments are linked together. Fragments are joined to each other by removing hydrogen atoms (See shaded boxes in step 1) at the points at which the fragments are to be linked. The bonding state of each atom is checked and, if necessary adjusted.

Step 3: the force field of the model is minimized to give the final structure.





An outline of the steps involved using INSIGHT II to produce a stick model of the structure of paracetamol

Molecular Dynamics

- Molecular dynamics is a molecular mechanics program designed to mimic the movement of atoms within a molecule.
- Molecular dynamics can be carried out o a molecule to generate different conformation which on energy minimization, give a range of stable conformation. Alternatively bonds can be rotated in a stepwise process to generate different conformation.
- Molecular dynamics can also be used to find minimum energy structures and conformational analysis.

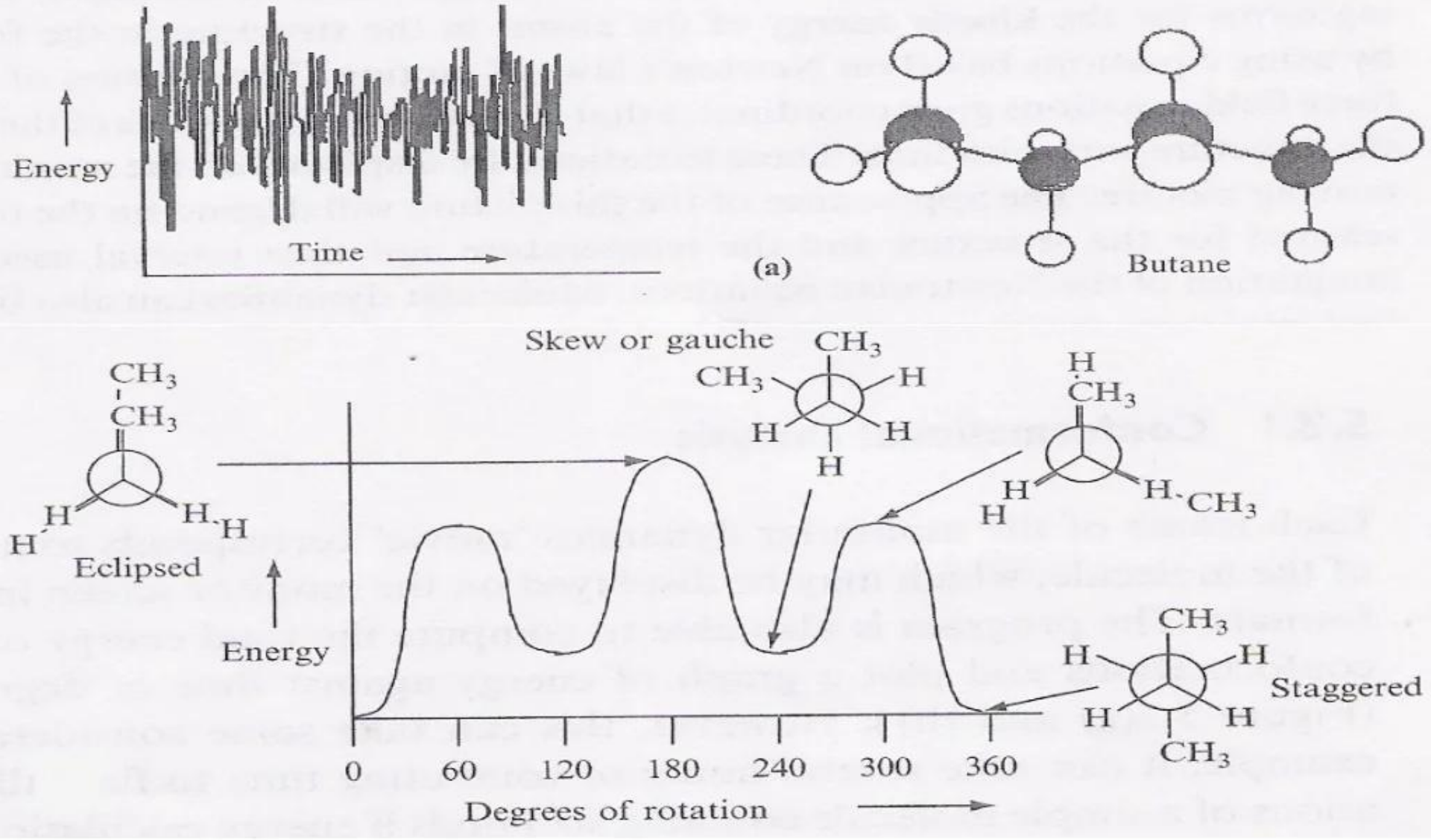


Fig-: Plot of change in the energy with the rotation about C_2-C_3 bond In butane

Quantum mechanics

- Quantum mechanics is based on arrangement of electrons of molecule and interaction of those electron with electron and nuclei of other molecule.
- It based on the realization that electron and all material exhibit wavelike properties.
- The Quantum mechanics based on finding solution to Schrödinger wave equation.

Schrödinger's Equation

- The Schrödinger equation is the basis of quantum mechanics and gives a complete description of the electronic structure of a molecule. If the equation could be fully solved all information of a molecule could be determined.

$$H\psi = E\psi$$

Where

H=Hamiltonian operator

Ψ =wave fuction

E =Energy system

To solve schrodinger equation was found to difficult. Hydrogen-total energy of hydrogen (E) can be described as the sum of kinetic energy and potential energy of its two component i.e. proton and electron.

- Schrödinger equation for this relationship can written,

$$H\psi=(k+v)=E\psi$$

Where as
K = kinetic energy
V = potential energy

- Describes both the wave and particle behavior of electrons.
  - The wave function is described by ψ while the particle behavior is represented by E.
  - In systems with more than one electron, the wave function is dependent on the position of the atoms; this makes it important to have an accurate geometric description of a system.

Quantum Mechanics Method

1. *Ab inito* method
2. Semiemperical method
3. Density functional theory

Ab Initio method

- *Ab initio* translated from Latin means from "first principles".
- This refers to the fact that no experimental data is used and computations are based on quantum mechanics.
- It derived directly from theoretical principle.

Different Levels of *Ab Initio* Calculations

1. Hartree-Fock (HF)

2. Density Functional Theory (DFT)

Hartree-Fock (HF)

- The simplest *ab initio* calculation.
- It based on Central field approximation.
- The major disadvantage of HF calculations is that electron correlation is not taken into consideration.

Density Functional Theory

- Considered an *ab initio* method, but different from other *ab initio* methods because the wave function is not used to describe a molecule.
- Density functional theory in which total energy is expressed in term of total electron density is used.
- DFT methods take less computational time than HF calculations and are considered more accurate.

Semi Empirical Method

- Semi-empirical quantum methods, represents a middle road between the mostly qualitative results available from molecular mechanics and the high computationally demanding quantitative results from *ab initio* methods.
- Semi empirical methods use experimental data to parameterize equations.
- Like the *ab initio* methods, a Hamiltonian and wave function are used.
- Less accurate than *ab initio* methods but also much faster.
- Capable of calculating transition states and excited states.

Choice of Method

The method of calculation based on what calculation needs to done and size of molecule.

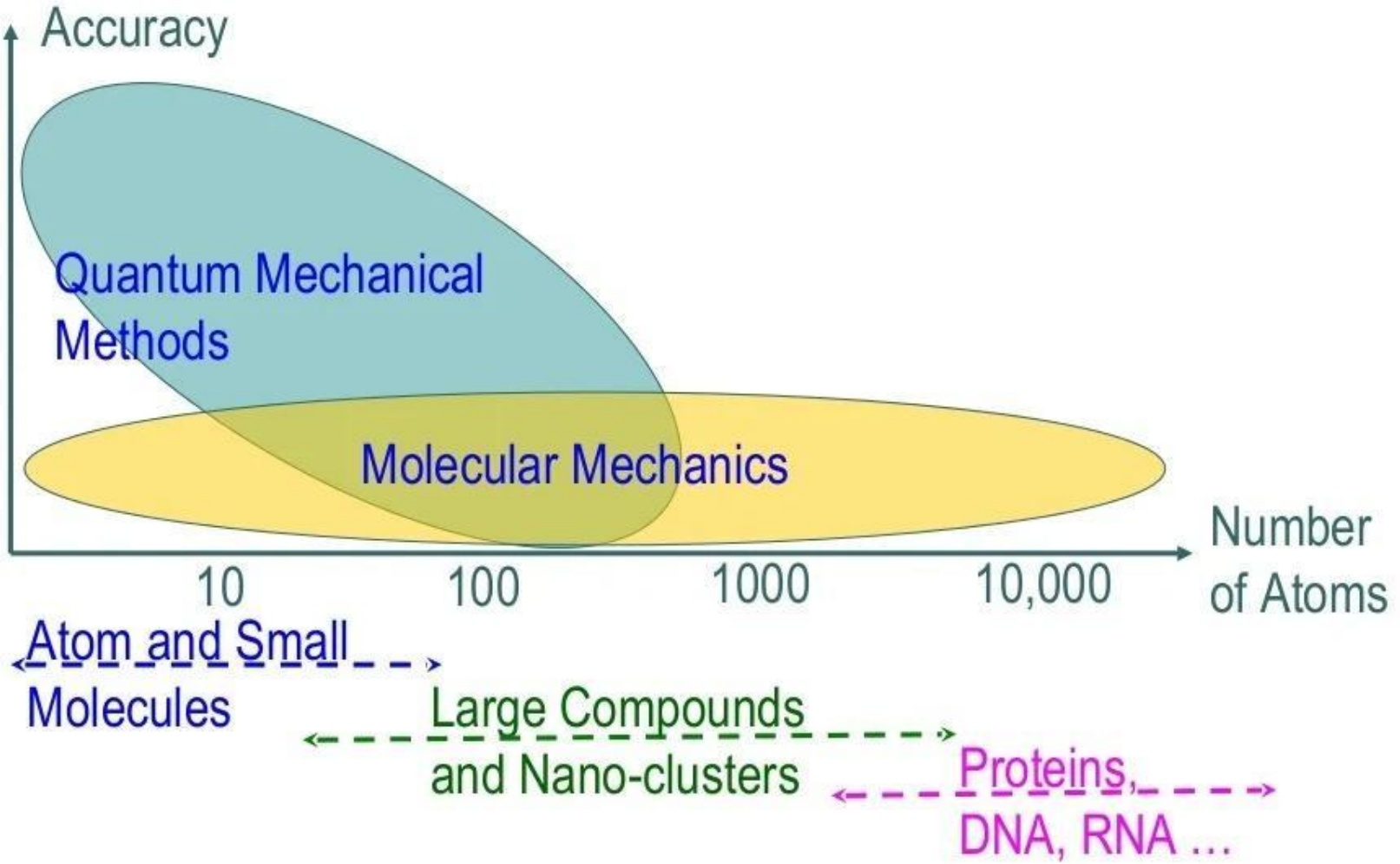
Molecular mechanics useful for

- Energy minimization
- Identifying stable conformation
- Energy calculation for specific conformations
- Studying molecular motion
- Studying different conformation.

Quantum mechanics method are suitable for calculating,

- Molecular orbital energies
- Heat formation for specific conformation
- Dipole moment
- Bond dissociation energy
- Transition-state geometries and energies

Computational Cost vs. Accuracy

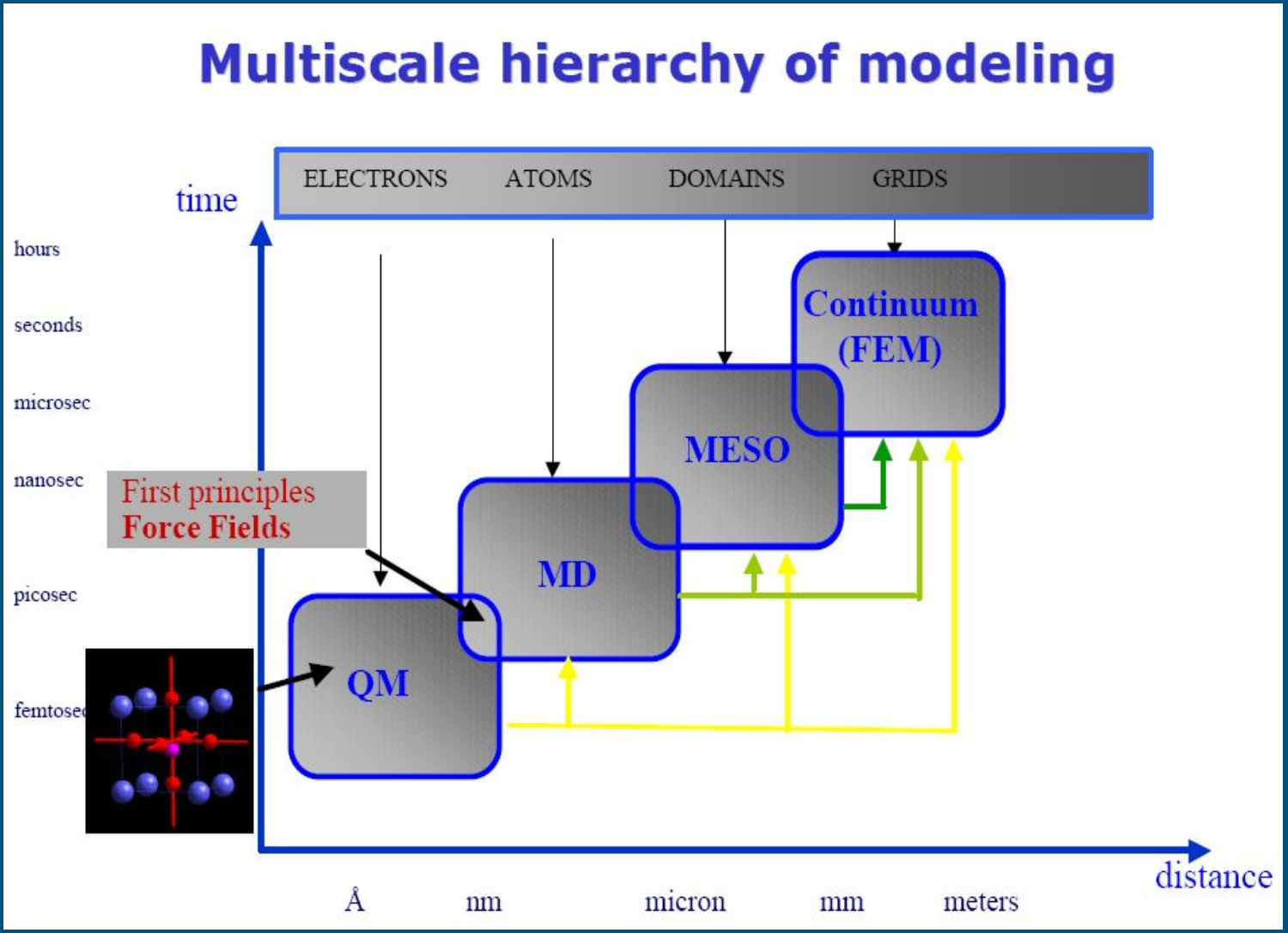


Quantum Mechanics vs. Molecular Mechanics

Quantum Mechanics	Molecular Mechanics
1. Correctly describes the Bond- breaking and Bond-forming	1. Does not properly describe the Bond-breaking and Bond-forming
2. Application limited to Hundreds of Atoms	2. Can treat more than 10,000 Atoms

Quantum mechanics	Molecular mechanics
1) More expensive	1) Less expensive
1) More time	2) Less time
1) More computing power	3) Less computing power
4) Used for small molecule	4) Used for large molecule
5) It can calculate electronic properties like electron density	5) It cannot calculate electronic properties like electron density
6) Application limited to hundreds of atom	6) Applicable to more than ten thousands of atoms

Multiscale Hierarchy of Modeling



How Big Systems Can We Deal with?

Assuming typical computing setup (number of CPUs, memory, disk space, etc.)

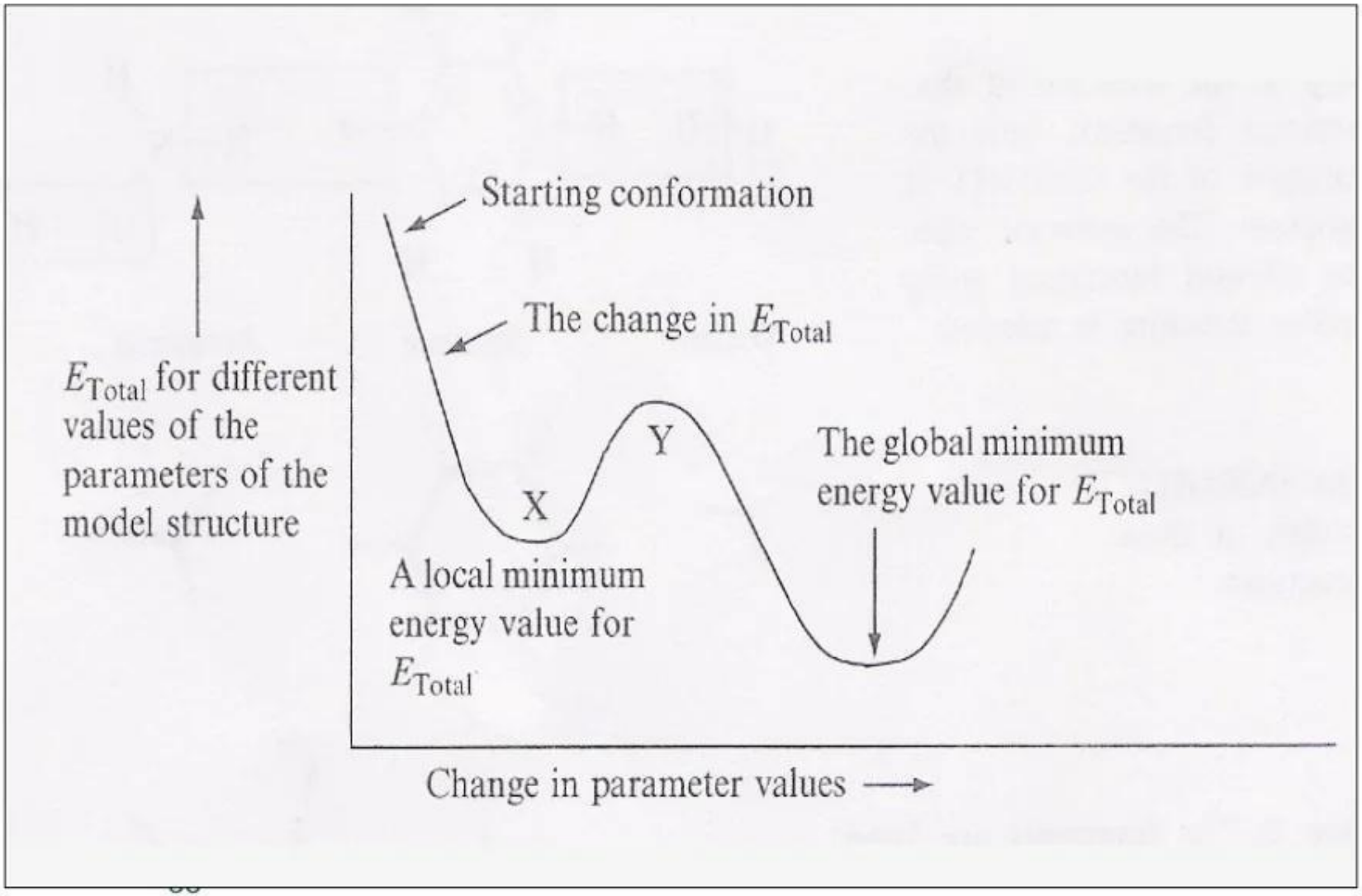
- Ab initio method: *~100 atoms*
- DFT method: *~1000 atoms*
- Semi-empirical method: *~10,000 atoms*
- MM/MD: *~100,000 atoms*

Energy minimization

- Energy minimization produces the nearest stable conformation to the structure presented and not necessarily the global conformation.

- Energy minimization involved alteration of bond length, bond angle, torsion angle and non-bonded interaction .

Energy minimization



Moleculer mechanics energy minimization

- Molecular mechanics is an approach of energy minimization that find stable, low energy conformation by changing the geometry of a structure.

Type of algorithms

1)Steepest Descent procedure

2)Conjugate gradient procedure

3)Newton –Raphson procedure

Selection of energy minimization algorithms

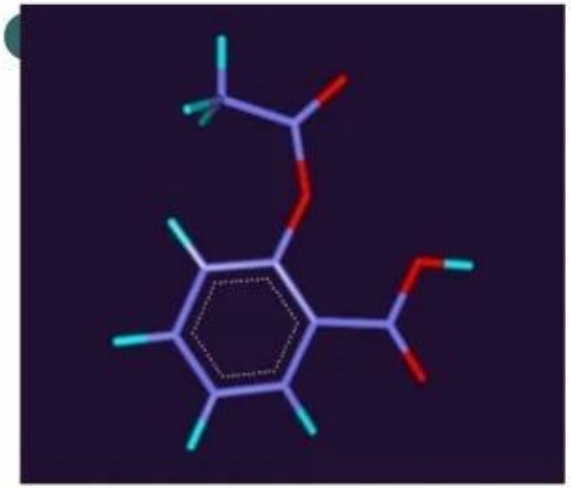
- The selection of energy minimization algorithms depends on size of system and current state of optimization.
- When molecule having larger than 200 atoms then conjugate gradient procedure.
- When molecule having less than 200 atoms then newton-raphson procedure.
- When molecule having larger than 10Kcal/mol/A then Steepest Descent.

Computer graphics

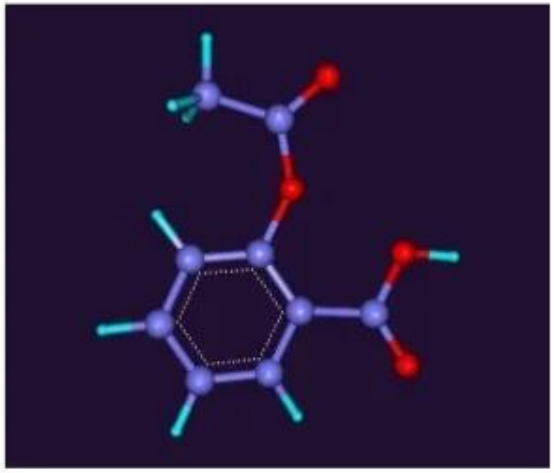
In Moleculer modeling, data produced are converted as visual image on a computer

Images displayed as,

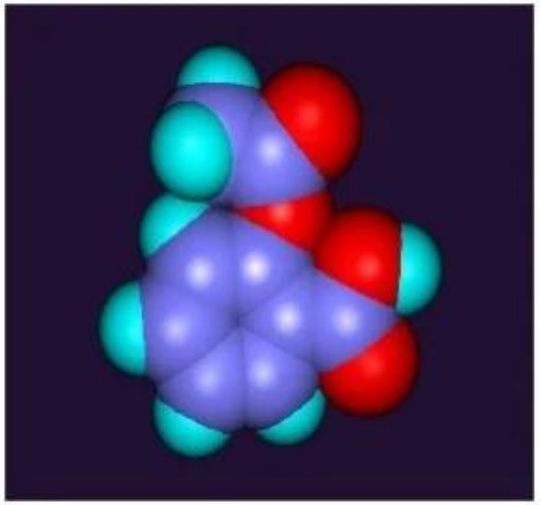
- Space fill model
- Ball and stick model
- CPK model
- Mesh model
- Ribbon model



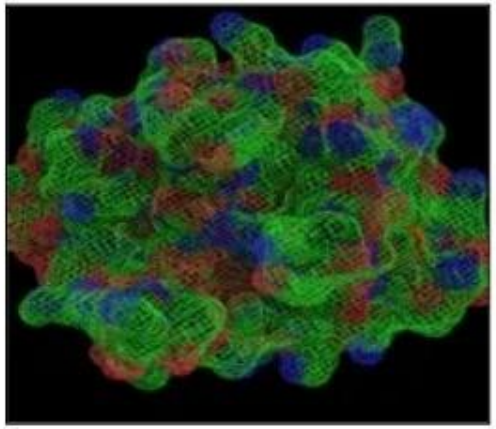
Stick model



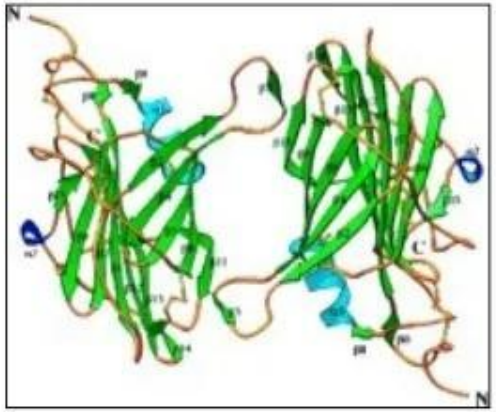
Ball and stick model



CPK model



Mesh representation of Aspirin

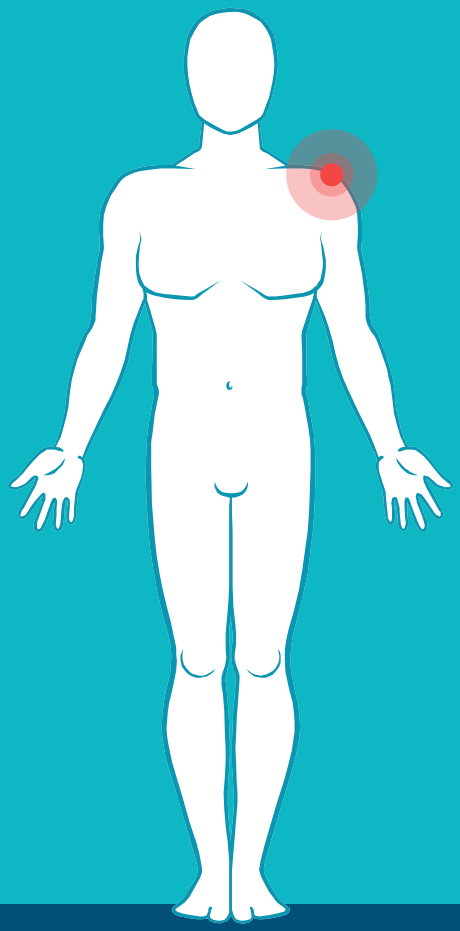


Ribbon representation of dihydrofolate reductase

Gene Duplication and Read Mapping

Week 7

Department of CSE, DIU



CONTENTS

1. Mutation
2. Gene Duplication
3. Read Mapping
   - Keyword Tree
   - Suffix Tree
   - Suffix Array
   - Burrows Wheeler Transform

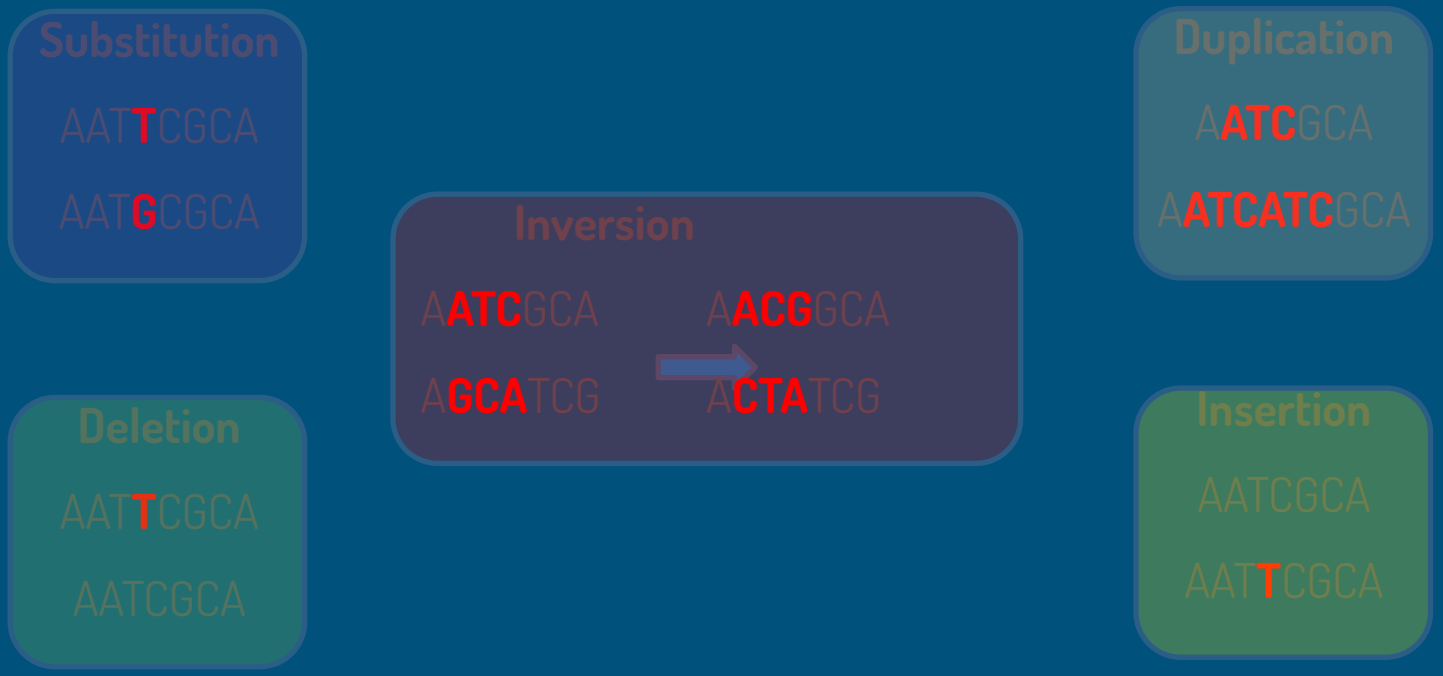
1. DNA Mutation

What and how mutation occurs, common forms

Mutation

DNA Mutation refers to sudden, random changes in DNA sequences which leads to different phenotypic expressions.

Common Mutation Types

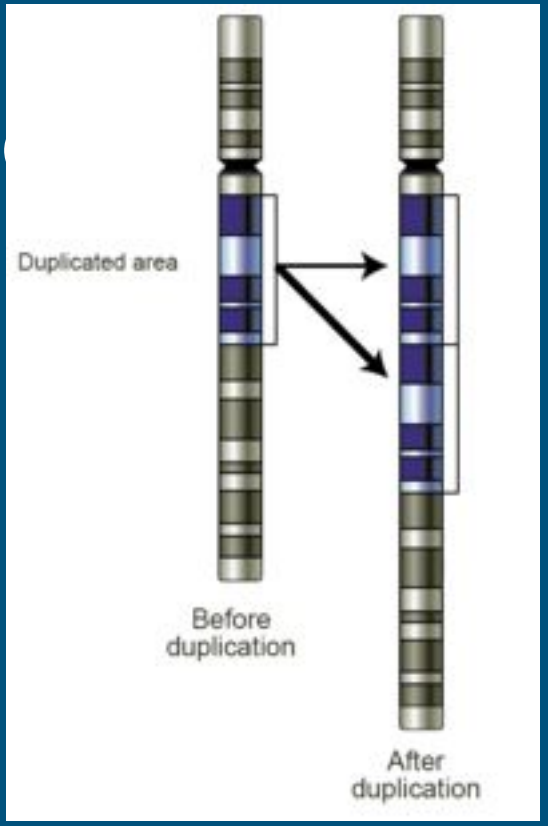


2. Gene Duplication

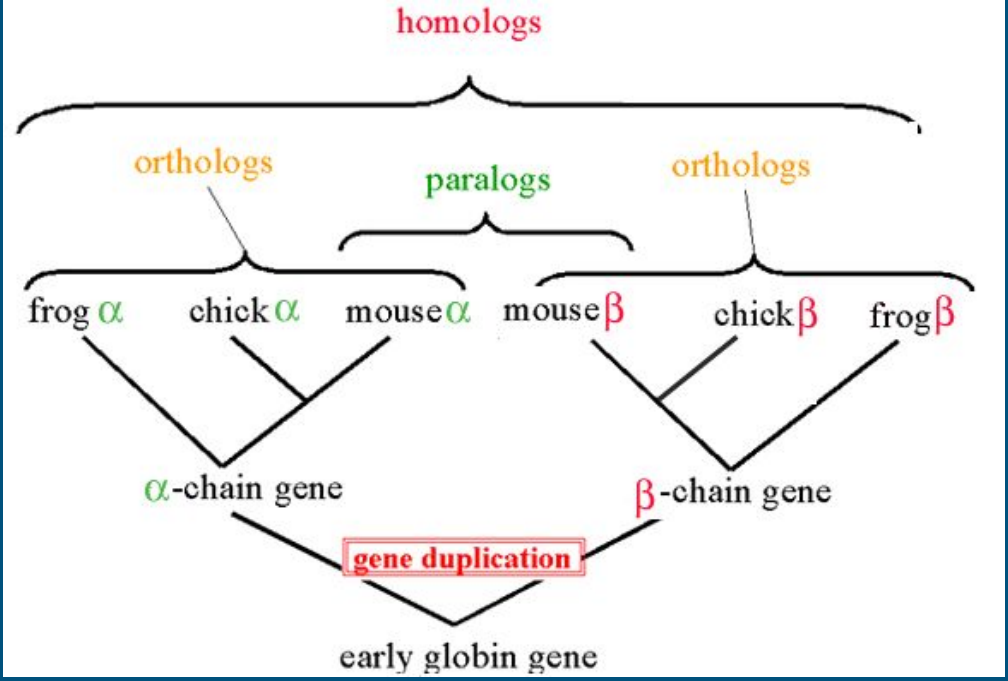
Duplication of Genes, Homolog, Ortholog, Paralogs

Gene Duplication

Gene duplication (or chromosomal duplication or gene amplification) is a major mechanism through which new genetic material is generated during molecular evolution. It can be defined as any duplication of a region of DNA that contains a gene.



Homolog, Ortholog, Paralog and Speciation



- Homolog – A gene related to a second gene by descent from a common ancestral DNA sequence
- Ortholog – Orthologs are genes in different species that evolved from a common ancestral gene by speciation*
- Paralog – Paralogs are genes related by duplication within a genome
- Speciation* – Speciation is the origin of a new species capable of making a living in a new way from the species from which it arose

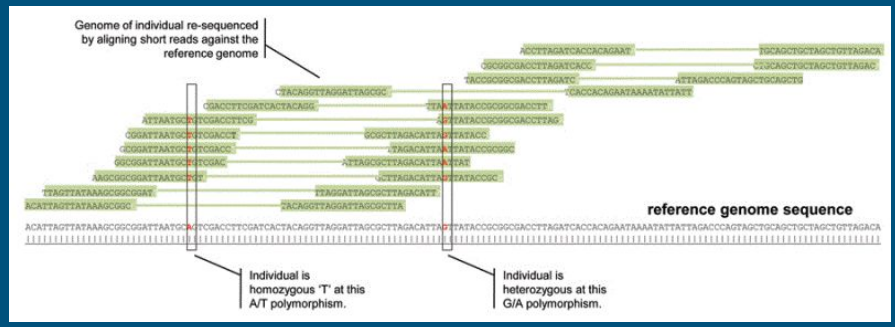
3. Read Mapping

Short Read Mapping, Genome Indexing

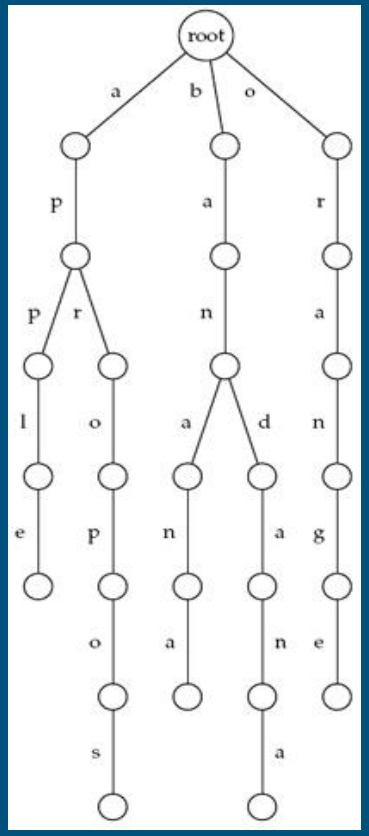
Read Mapping

Mapping refers to the process of aligning short reads to and finding the starting position in a reference sequence (typically Genome).

Short read generally are reads with a length of 30-350 base pairs.



Genome Indexing (Keyword Tree)



- Stores a set of keywords in a rooted labeled tree.
- Each edge is labeled with a letter from an alphabet.
- Any two edges coming out of the same vertex have distinct labels.
- Every keyword stored can be spelled on a path from root to some leaf.
- Furthermore, every path from root to leaf gives a keyword.

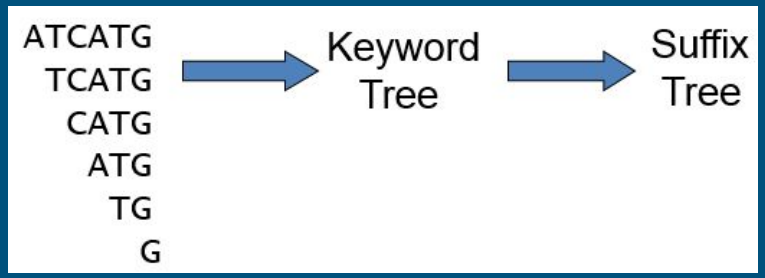
Keywords

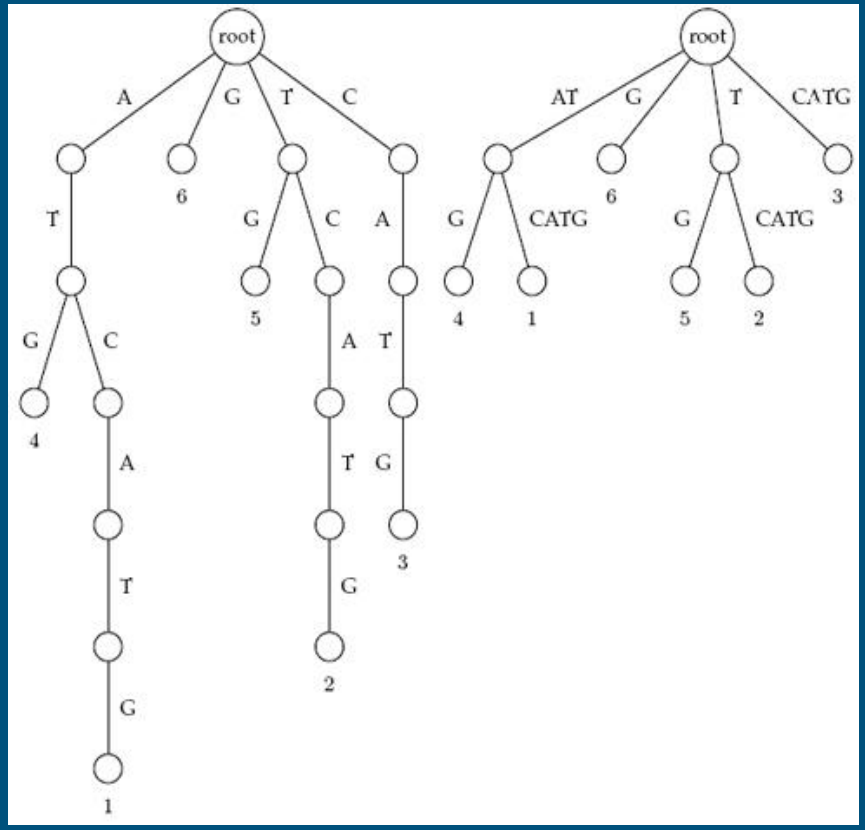
- Apple
- Apropos
- Banana
- Bandana
- Orange

Genome Indexing (Suffix Tree)

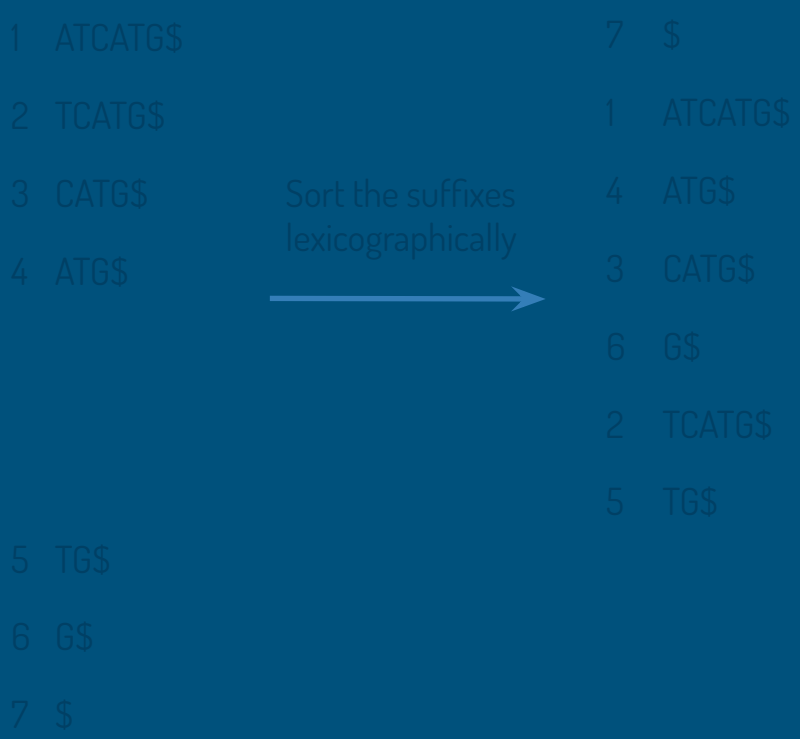
- **Similar to Keyword Tree**
- Suffixes of the text are keywords
- Edges that form paths are collapsed
- Each edge is labeled with a substring of the text
- All internal edges have at least two outgoing edges.
- Leaves are labeled by the index of the pattern.

Suffix tree of ATCATG





Genome Indexing (Suffix Array)

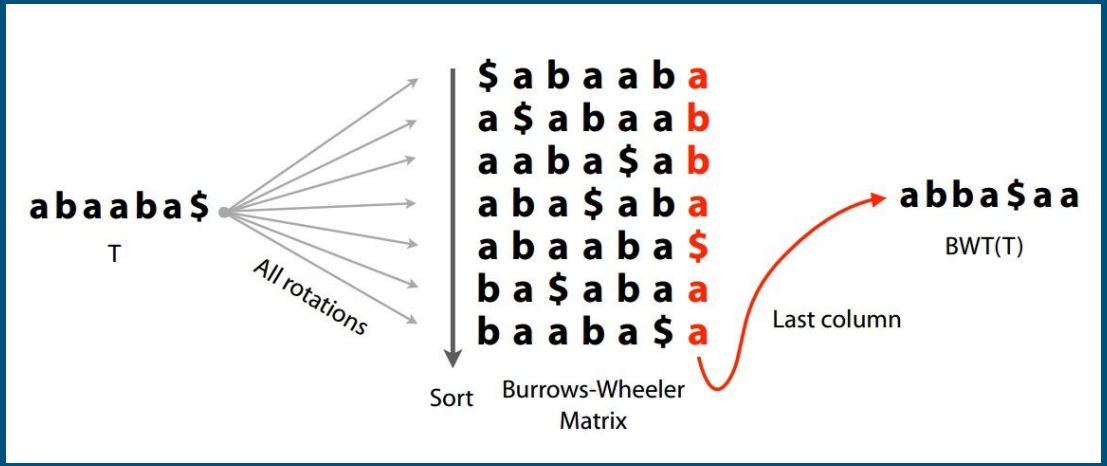


- **More space efficient than suffix tree**
- Suffix tree index for human genome is about 47 GB
- Lexicographically sort all the suffixes
- Store the starting indices of the suffixes along with the original string

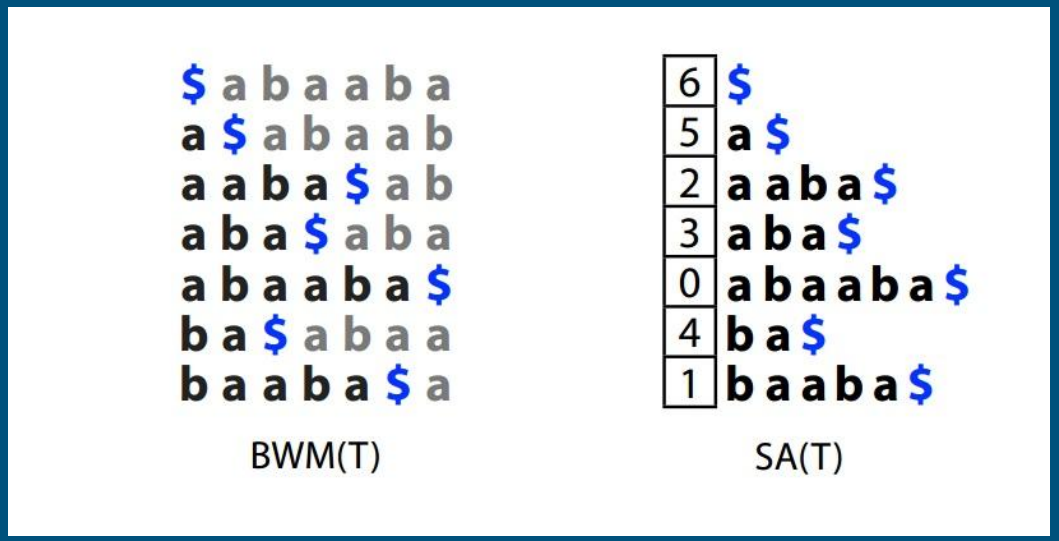
Generate Suffix Array of ATCATG

Genome Indexing (Burrows Wheeler Transform)

- Given Sequence – **abaaba**
- Add **$** as ending notation – **abaaba$**
- By Shifting each alphabet to the right once, generate all the rotations
- Lexicographically Sort all the rotations
- The very last column will be denoted as BWT (T)



Genome Indexing (Burrows Wheeler Transform)

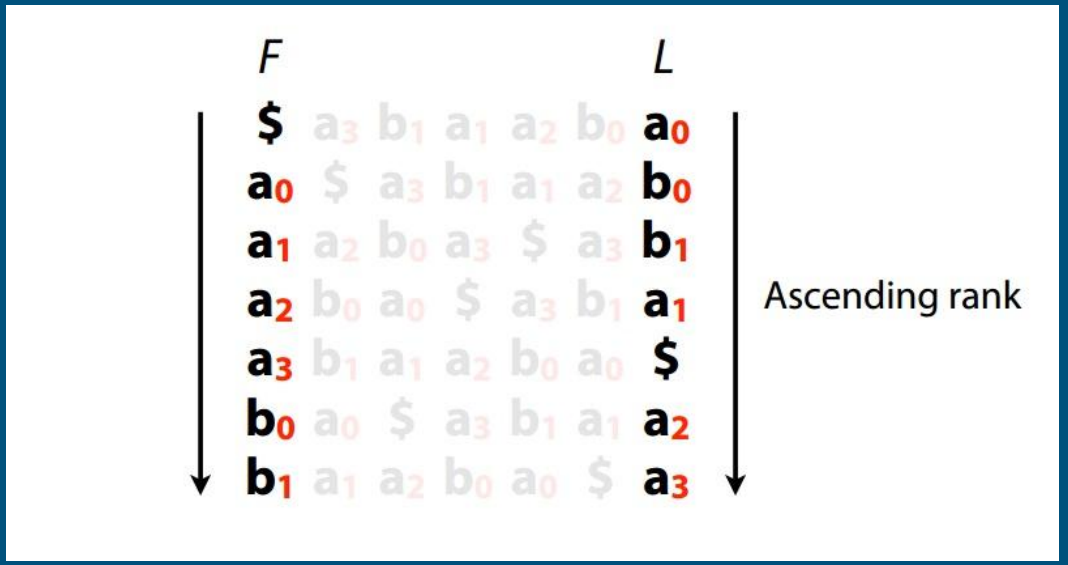


- Given Sequence – **abaaba**
- Add **$** as ending notation – **abaaba$**
- Lexicographically sorted all rotations will generate BWT Matrix which will be denoted as BWM (T)
- Suffix Array generated from all the rotations will be called SA (T)
- BWM can be derived from any given BWT (T)

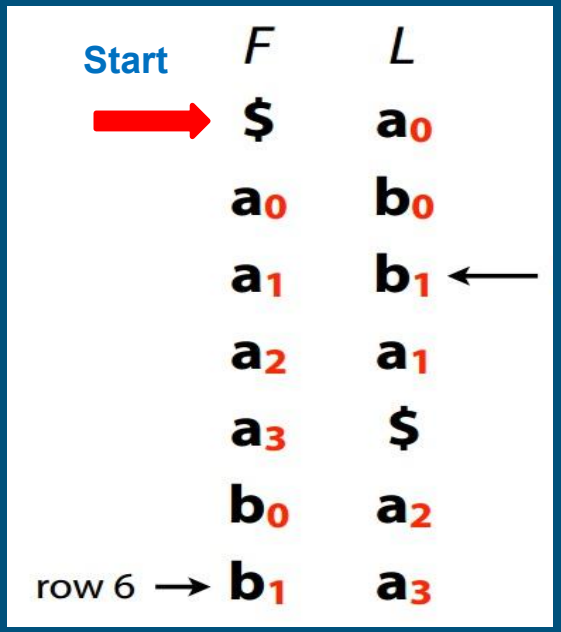
Genome Indexing (Burrows Wheeler Transform)

LF (Last to First) Mapping

- Generate Burrows Wheeler Matrix for a given sequence
- Assign numbers to distinguish same characters
- Assign the numbers in a ascending manner for each character



Genome Indexing (Burrows Wheeler Transform)



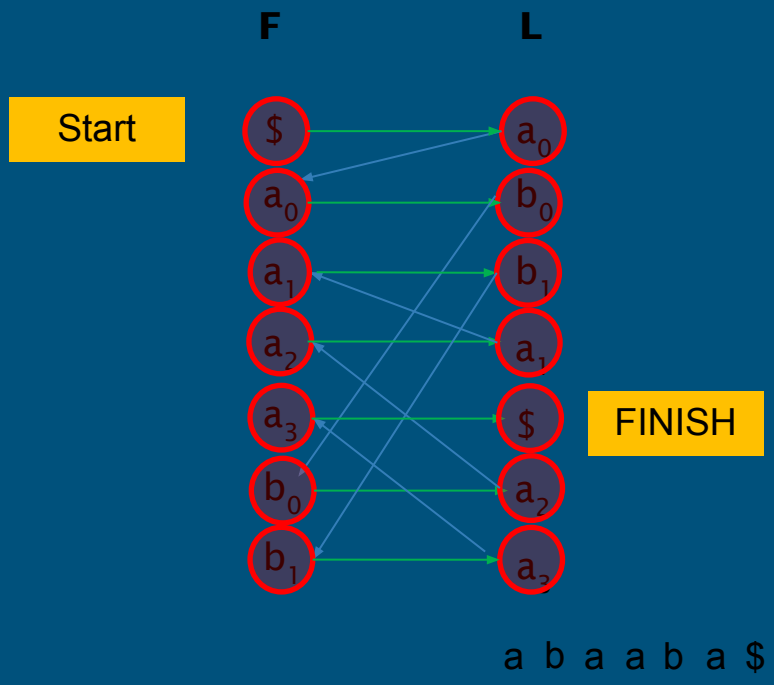
Find out the row starting with b1 using LF Mapping

1. Start from the row containing **$** in the First Column
2. Find out what's in Last Column of that row (here its a_0)
3. Compare it with query **(b_1)**
4. If MATCH, then
   - Find **b1** in First Column
   - Print row number
   - Terminate
5. If No MATCH, then
   - Find the row with that element in the First column
   - Go to Step 2 and Repeat

Genome Indexing (Burrows Wheeler Transform)

Find Original Gene using LF Mapping if BWT (T) is Given

1. Original Gene = **abaaba** (Not Given)
2. Given BWT (T) = **abba$aa**
3. Store it as Last Column
4. Draw the First Column by sorting the elements of Last Column Lexicographically
5. Assign numbers to distinguish characters in an ascending manner
6. Start LF Mapping from Starting Element ($)
7. For each element found in the **LAST** column, write it from right to left



Whales and Dolphins

Their ancestors had back legs once, they could walk

Birds came from Dinosaurs

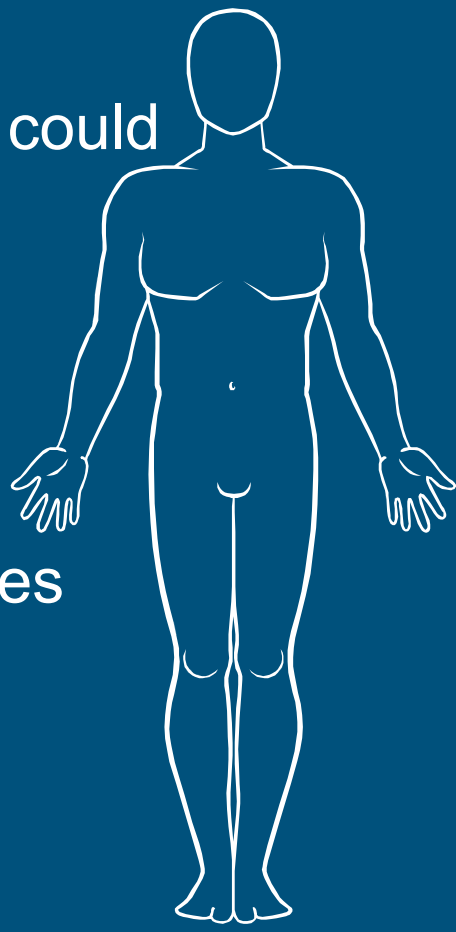
And they both descended from Reptiles

Humans have tails

While they are inside the womb! It dissolves eventually.

Bacterium

All livings beings can be traced back to a bacterium



Database Searching (FASTA)

Lecture – 8

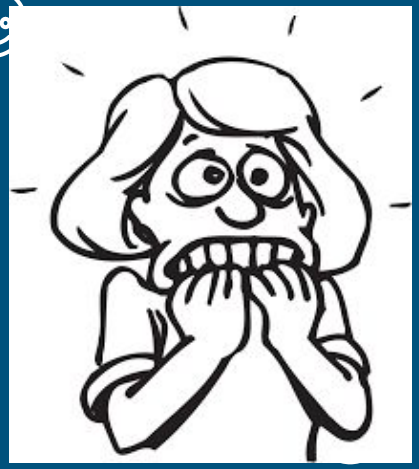
Department of CSE, DIU

CONTENTS

1. TP, TN, FP, FN

True Positive, True Negative, False Positive, False Negative

A patient fears that he has Cancer
&
Goes to the doctor for Diagnosis



Possible Scenarios

True Positive

Patient really had cancer
&
Diagnosis came Positive

True Negative

Patient didn't have cancer
&
Diagnosis came Negative

False Positive

Patient didn't have cancer
&
Diagnosis came Positive

False Negative

Patient really had cancer
&
Diagnosis came negative

2. Selectivity and Sensitivity

We will learn about calculating selectivity and sensitivity

Selectivity & Sensitivity

$$\text{Sensitivity} = \frac{True\ Positive}{All\ Positive}$$

$$\text{Selectivity} = \frac{True\ Negative}{All\ Negative}$$

Worked Out Example (Sensitivity)

Dataset	
A	G
C	T
G	T
G	C
A	G
C	G

Search Character = C
Expected = CCC
Outcome = ACC

$$\text{Sensitivity} = \frac{True\ Positive}{All\ Positive}$$

- Suppose we are searching the character C in entire database
- Each time we encounter a C we should print C
- So the final output of search should be = CCC as there are 3 Cs in the entire dataset. But the outcome is ACC
- So, **All Positive = 3** (as there are 3 Cs in the whole dataset and we are looking for C only)
- **True Positive** = number of Cs in the outcome ACC = 2
- Sensitivity = $\frac{2}{3}$

Worked Out Example (Selectivity)

Dataset

A	G
C	T
G	T
G	C
A	G
C	G

Search Character = C
Expected = CCC
Outcome = ACC

▷ $\text{Selectivity} = \frac{True\ Negative}{All\ Negative}$

- Suppose we are searching the character C in entire database
- Each time we encounter a C, we should print C
- So the final output of search should be = CCC as there are 3 Cs in the entire dataset. But the outcome is ACC
- So, **All Negative = 9** (Number of entries in the dataset that is not C)
- **True Negative** = number of entries in the outcome ACC that is not C = 1
- Sensitivity = $\frac{1}{9}$

3. Hash Table Used in FASTA

Hash Table Algorithm

Given Data

Query Sequence: JUSTICELEAGUE
Target Sequence: LEAGUEOFASSASINS
Value of K : 1

Step 1 : Build Query Table

1	2	3	4	5	6	7	8	9	10	11	12	13
J	U	S	T	I	C	E	L	E	A	G	U	E

Step 2: Hash Table for Query Sequence

Write all the distinct characters appeared in the Query Sequence Lexicographically and then, beneath that, write the number of the position in which that letter appeared. There can be multiple occurrences.

A	C	E	G	I	J	L	S	T	U
10	6	7	11	5	1	8	3	4	2
		9							12
		13							

Step 3 : Build Target Table

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
L	E	A	G	U	E	O	F	A	S	S	A	S	I	N	S

Step 4 : Import the Hash Table for Query Sequence

A	C	E	G	I	J	L	S	T	U
10	6	7 9 13	11	5	1	8	3	4	2 12

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
L	E	A	G	U	E	O	F	A	S	S	A	S	I	N	S

Step 5 : Build the Extended Target Table based on Hash Table

A	C	E	G	I	J	L	S	T	U
10	6	7 9 13	11	5	1	8	3	4	2 12

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
L	E	A	G	U	E	O	F	A	S	S	A	S	I	N	S
7	5 7 11	7	7	-3 7	1 3 7			1	-7	-8	-2	-10	-9		-13

Entry in Extended Row = Position of the Letter in Hash Table – Position of the Letter in Extended Target Table
Example:

- For L, in Extended Target Table, Entry is 7 (8-1) .
- Similarly For E, the entries are 5 (7-2), 7 (9-2) and 11 (13-2).

Step 5 : Build Offset Table

Draw a table from the minimum to the maximum entry of the extended target table. Then beneath each entry number, write down number of times that entry occurred in extended target table. For example, the entry 7 Occurred 6 times and the entry 1 occurred 2 times.

-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11
1			1	1	1	1				1	1			2		1		**1**		6				**1**

Step 6: Build Pre-Final Table

Start both Query and Target sequence from 0 position.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
J	U	S	T	I	C	E	L	E	A	G	U	E			
L	E	A	G	U	E	O	F	A	S	S	A	S	I	N	S

Step 7 : Build Final Table

- Find out the entry number from the offset table, that occurred maximum number of times (Here 7, which occurred 6 times).
- After that, add that entry number with the previous starting position of target sequence to get the new starting Position of Target Sequence (Previous starting position = 0, Then new starting position of target seq becomes 0 + 7 = 7).

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
J	U	S	T	I	C	E	L	E	A	G	U	E										
							L	E	A	G	U	E	O	F	A	S	S	A	S	I	N	S


Thanks!