Introduction to Biology and Chemistry for Computation

Presented by:
Tanim Ahmed
B.Sc. in CSE,KUET
Lecturer(CSE),DIU
tanimahmed.cse@diu.edu.bd

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.

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


QUESTIONS

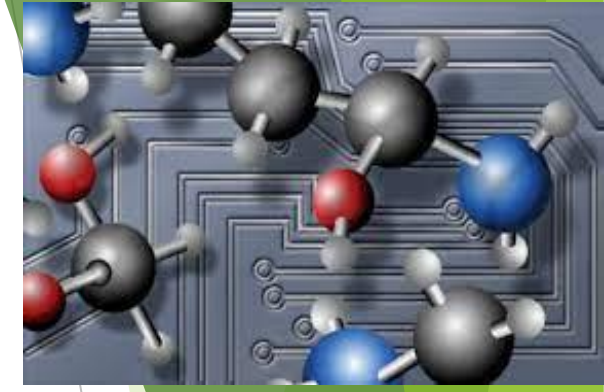
Lecture 1.2

Role of chemistry in computer science & engineering

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)

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

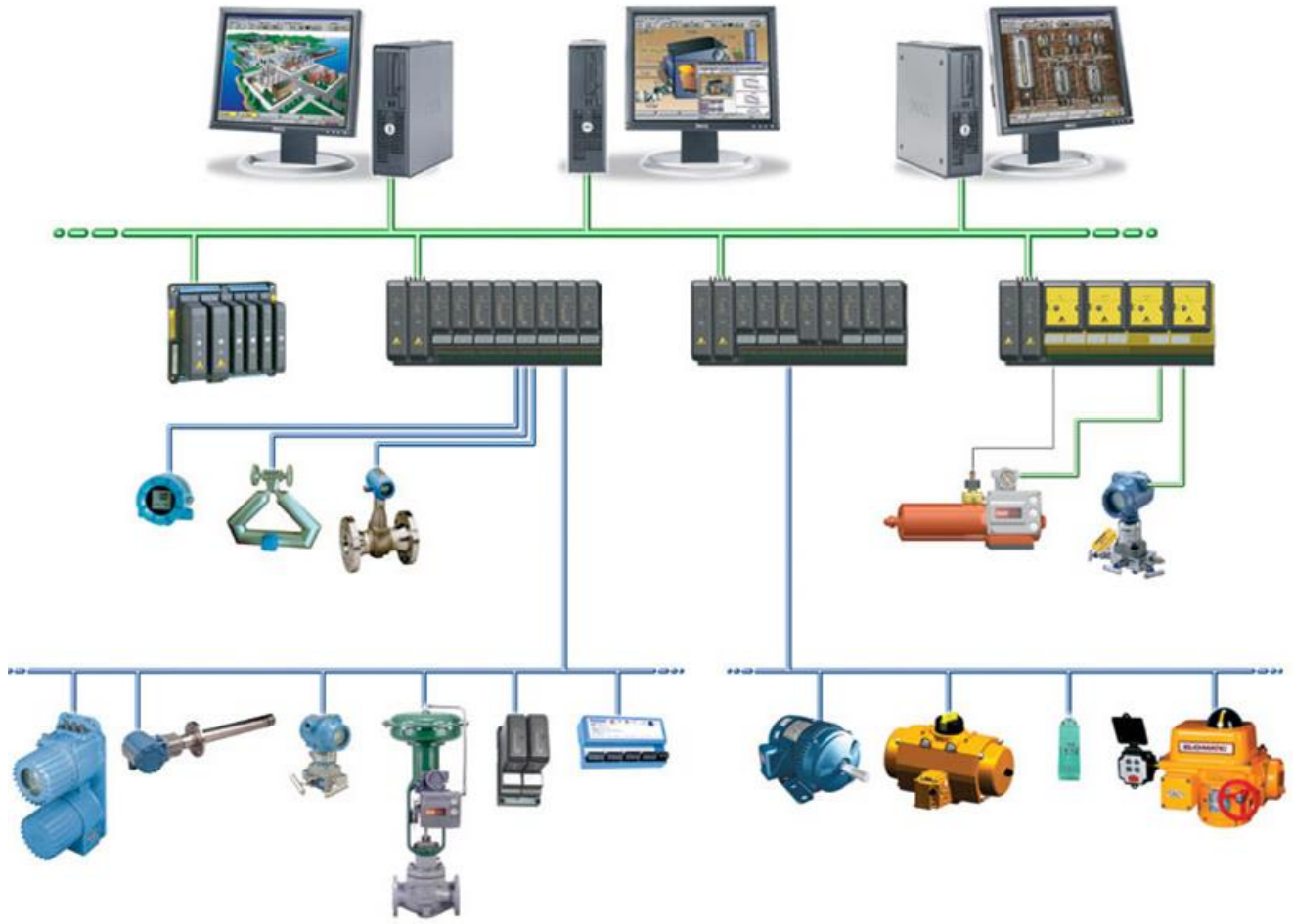
- Some others areas are :

- Chemical plants
- Petrochemical (oil) and refineries
- Pulp and paper mills (see also: quality control system QCS)
- 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
- Agriculture applications

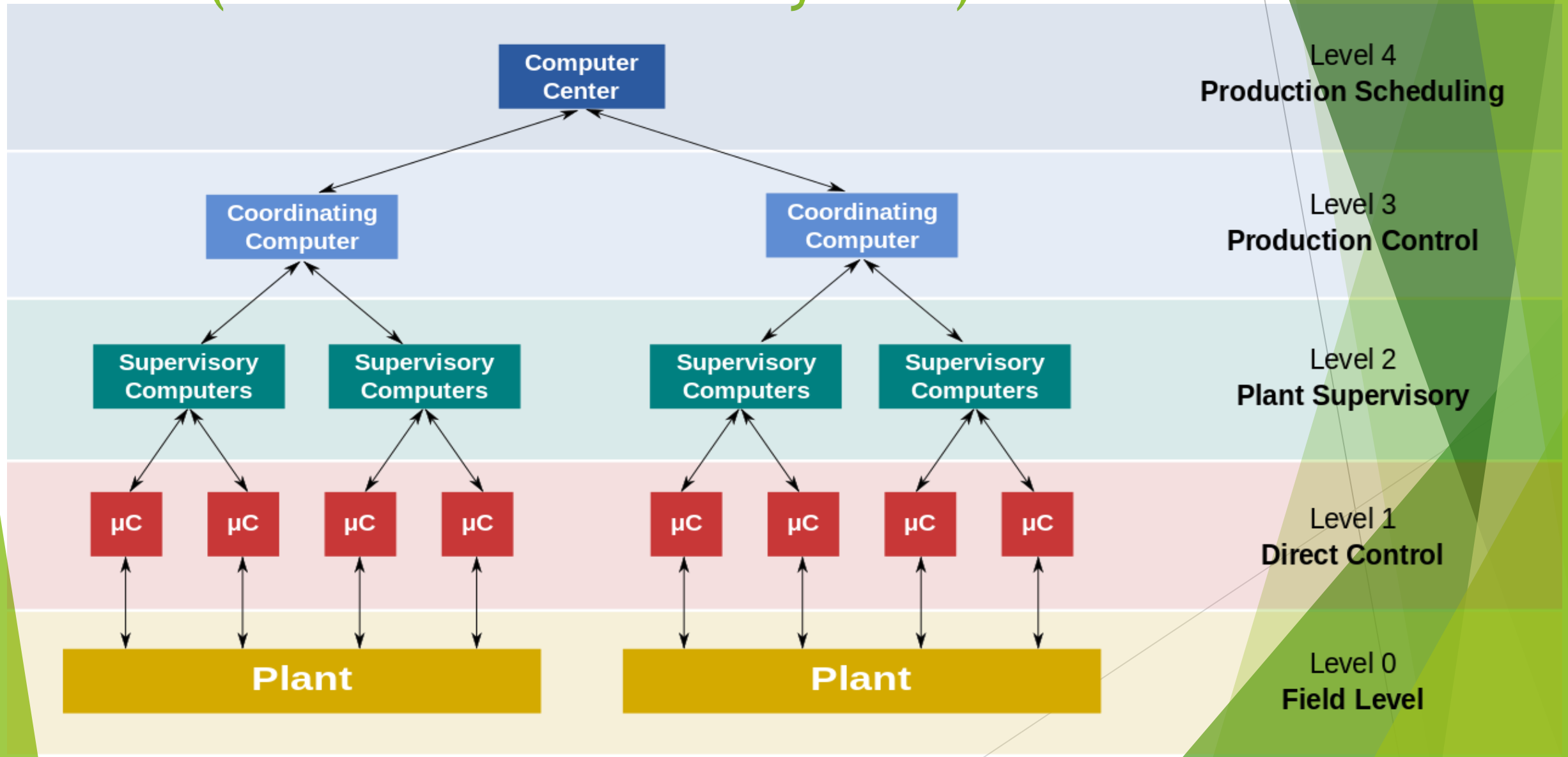
DCS (distributed control system)

- 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)

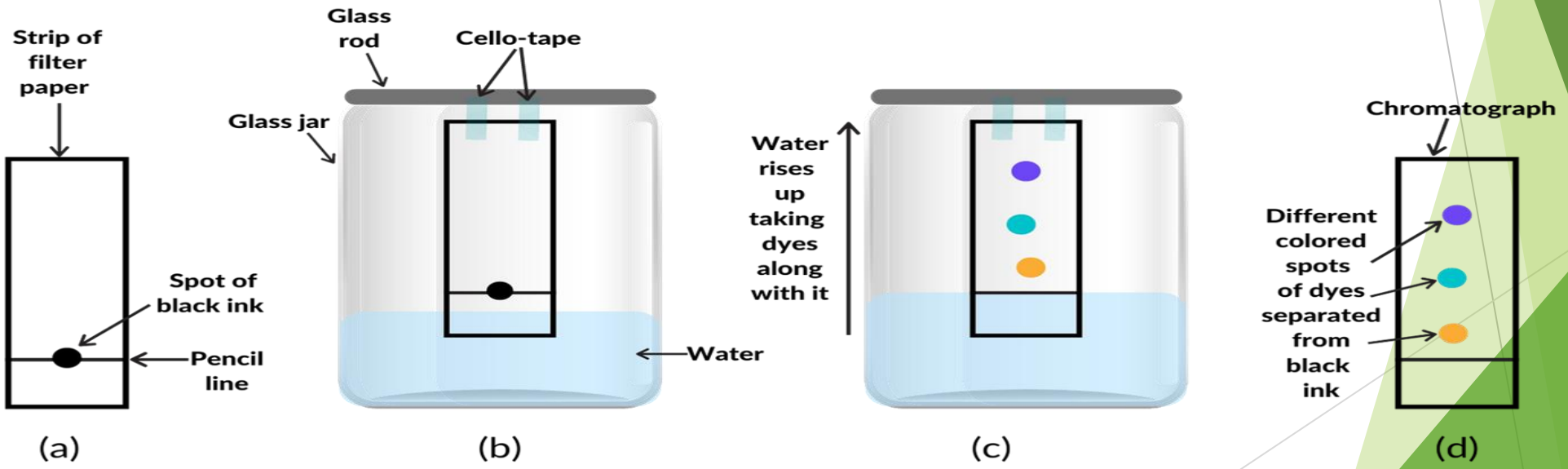


Chromatography

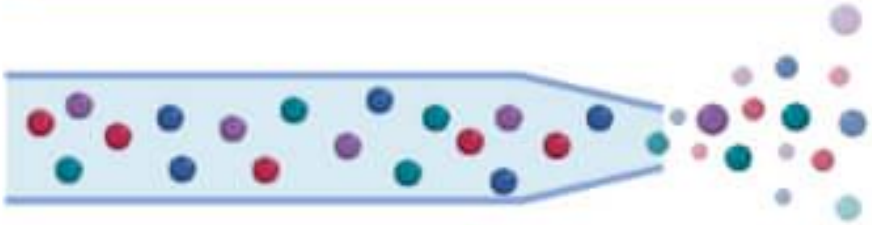
- Chromatography is **a process for separating components of a mixture**
- **Now a days computer aided systems are being used for chromatography.** Chromatography is used to separate a mixture of sample causing them to separate. Using a computer to analyze the time taken for a compound to be detected, 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.

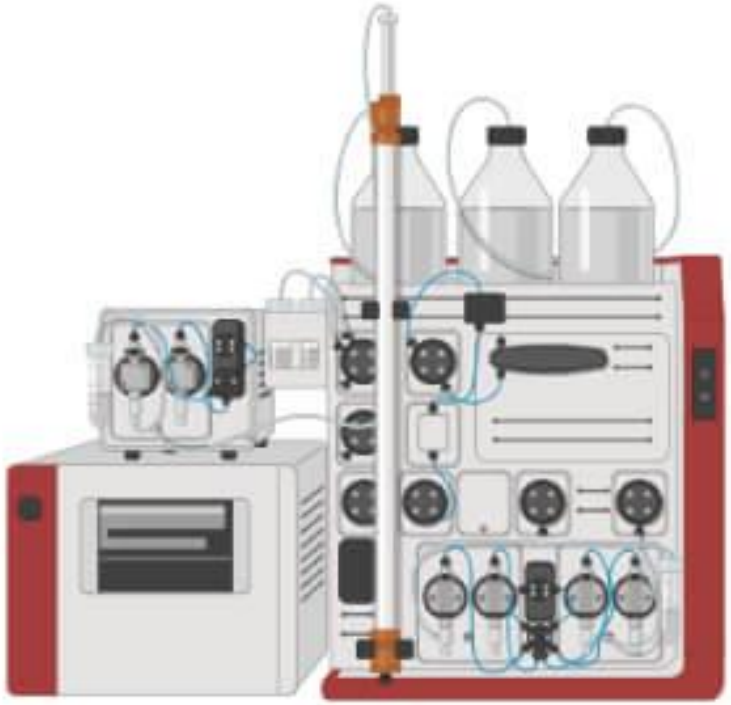
teachoo

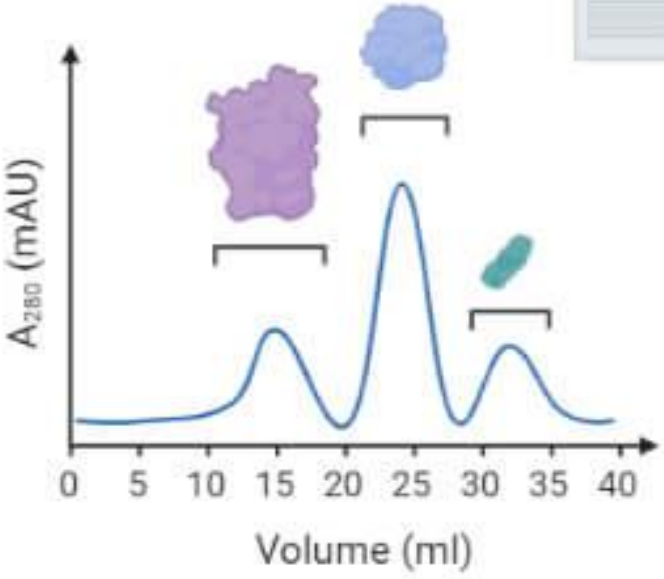
Chromatography



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

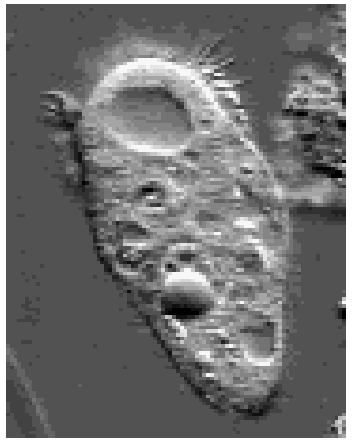
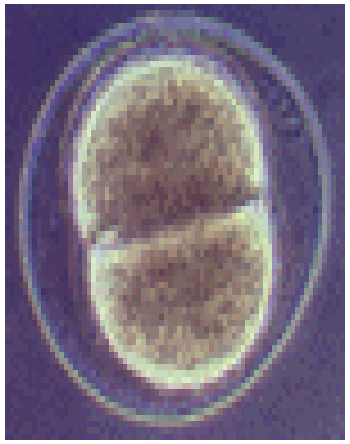
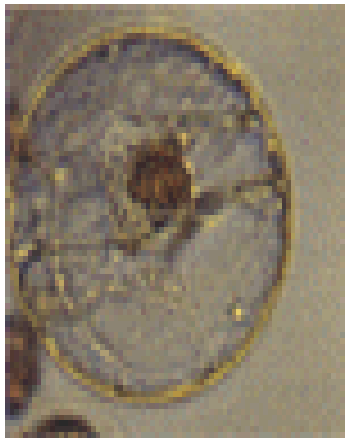
Molecular and Cellular Biology

Lecture – 2

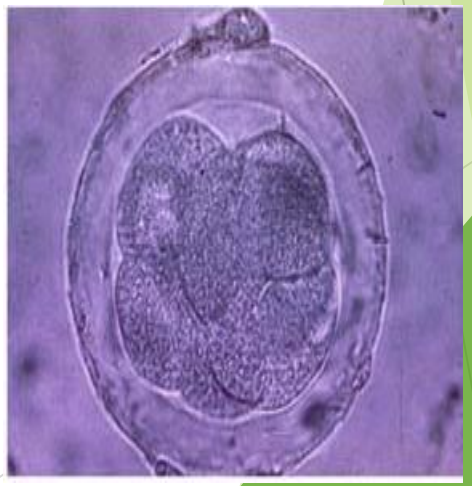
Department of CSE, DIU

CONTENTS

What is Life made of?



1. Cell

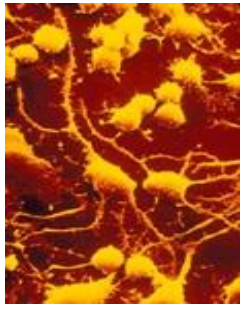
Let’s learn about Eukaryotes and Prokaryotes

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

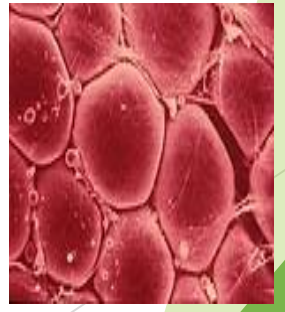
Blood

Bone



Nerve

Muscle

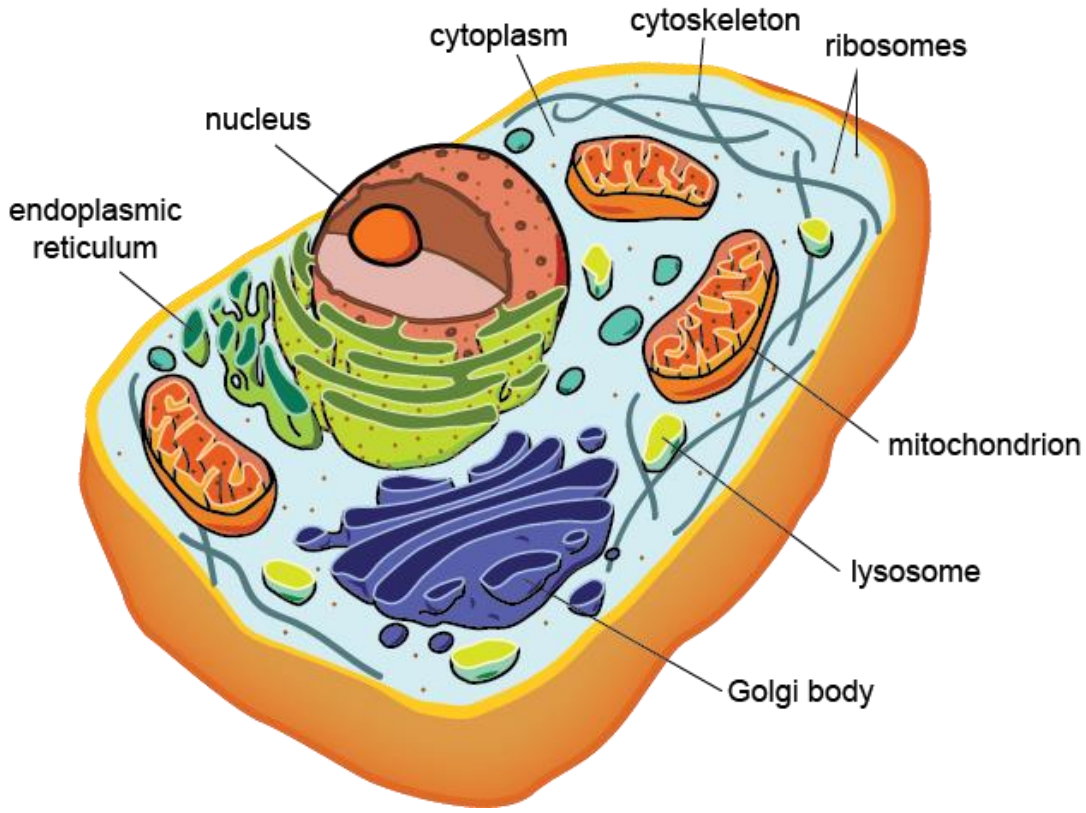


Fat

2 types of Cells

1. Eukaryotic Cells
2. Prokaryotic Cells

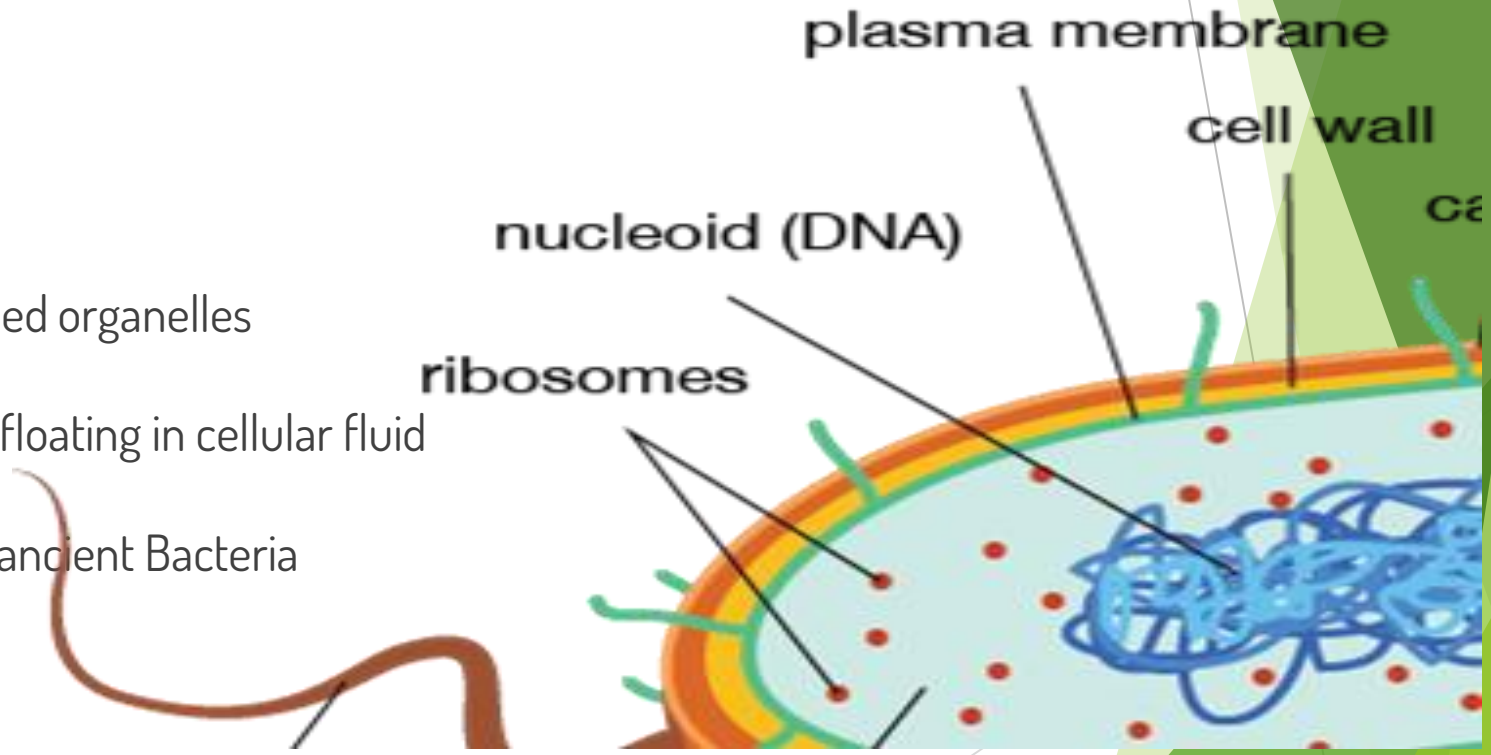
Eukaryotic Cells



- 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



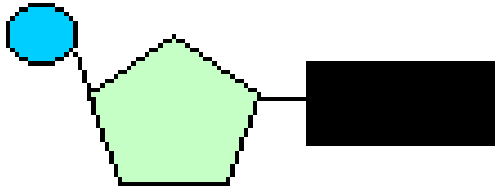
3. Nucleic Acid

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

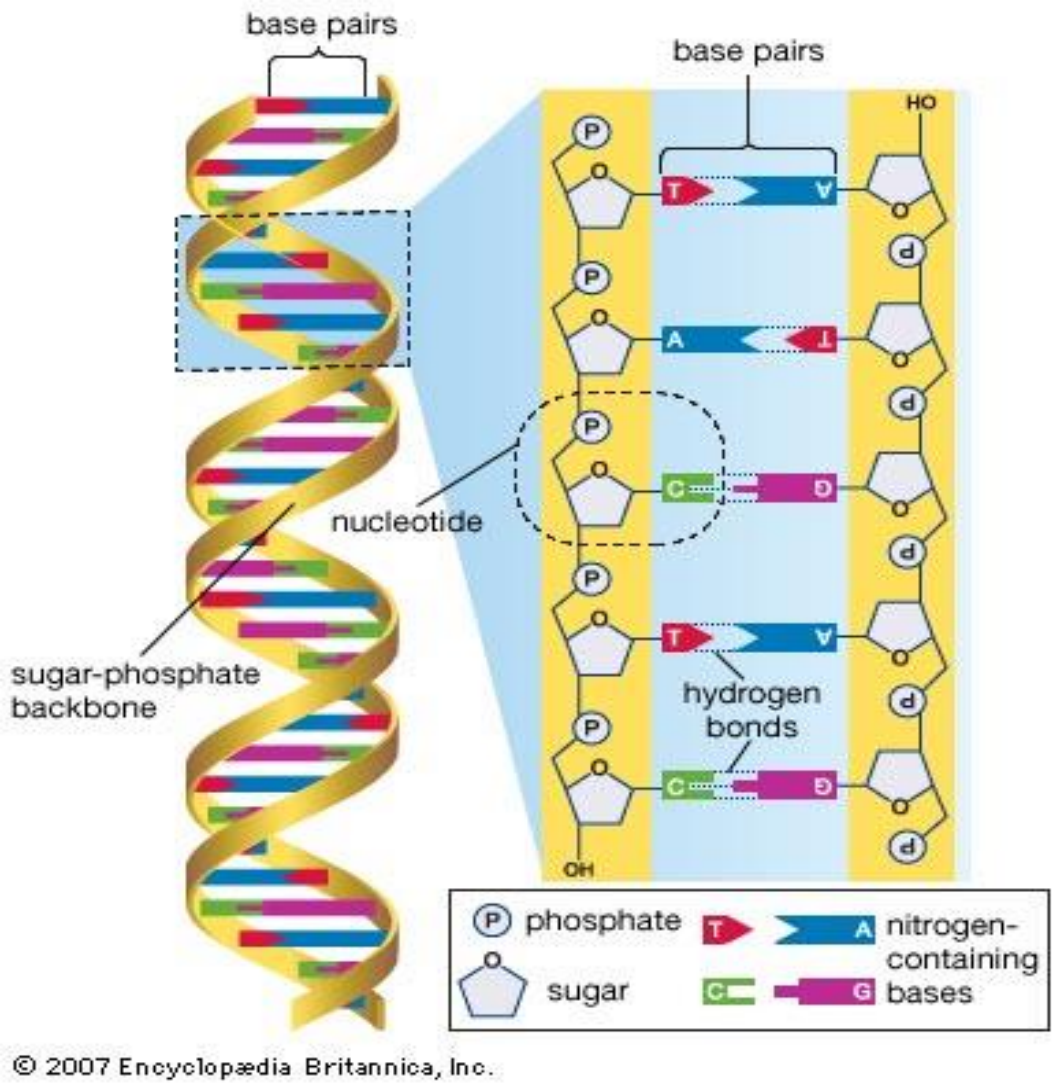
- Adenine (A),
- Guanine (G)
- Cytosine (C)
- Thymine (T)
- Uracil (U)

3.1 DeoxyriboNucleic Acid (DNA)

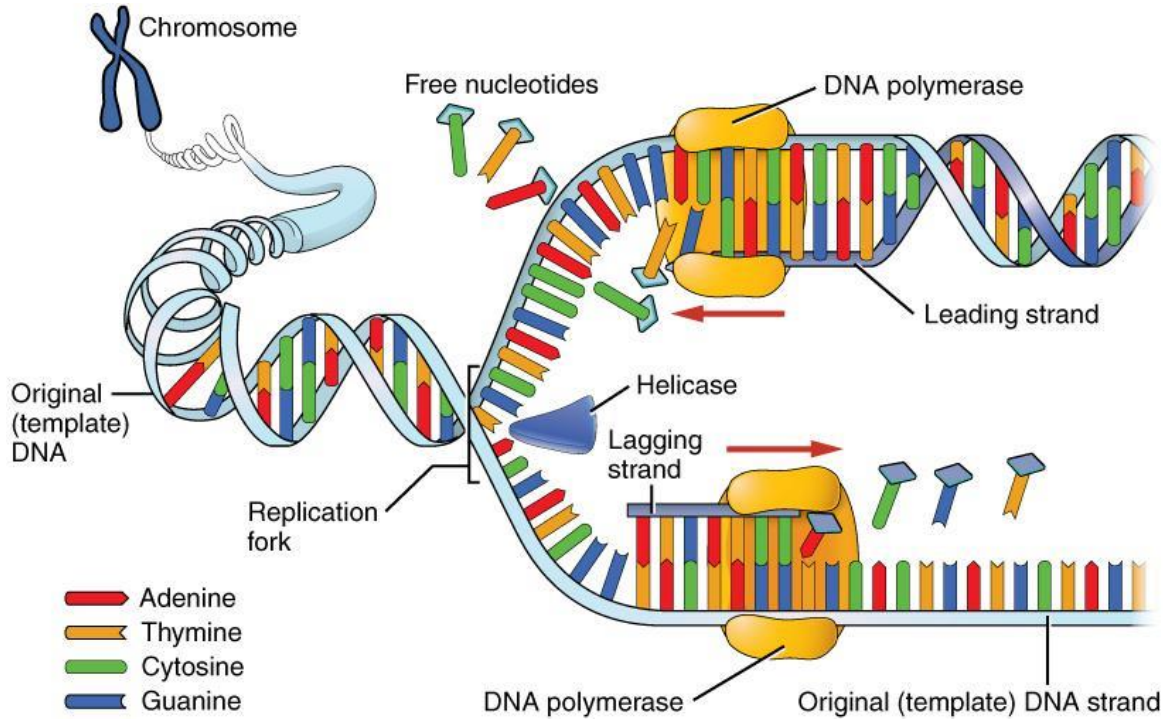
Carrier of genetic instructions

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

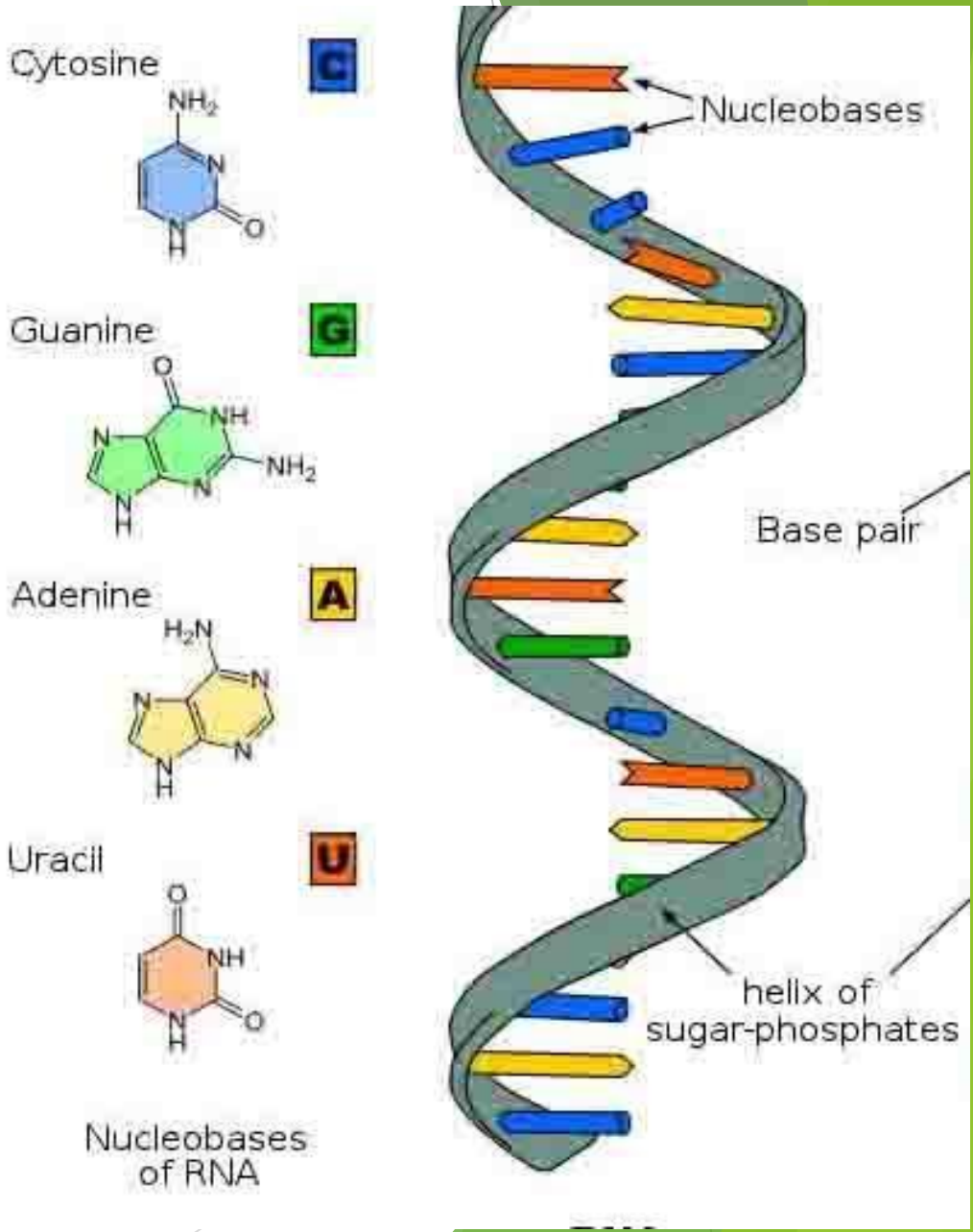
*** DNA Replication is Semi-Conservative, because, in new sets of DNA, one strand is newly created but the other strand comes from the ancestor.**

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

Double Stranded RNA

Catalyze complementary strands like DNA. Induces gene expression.

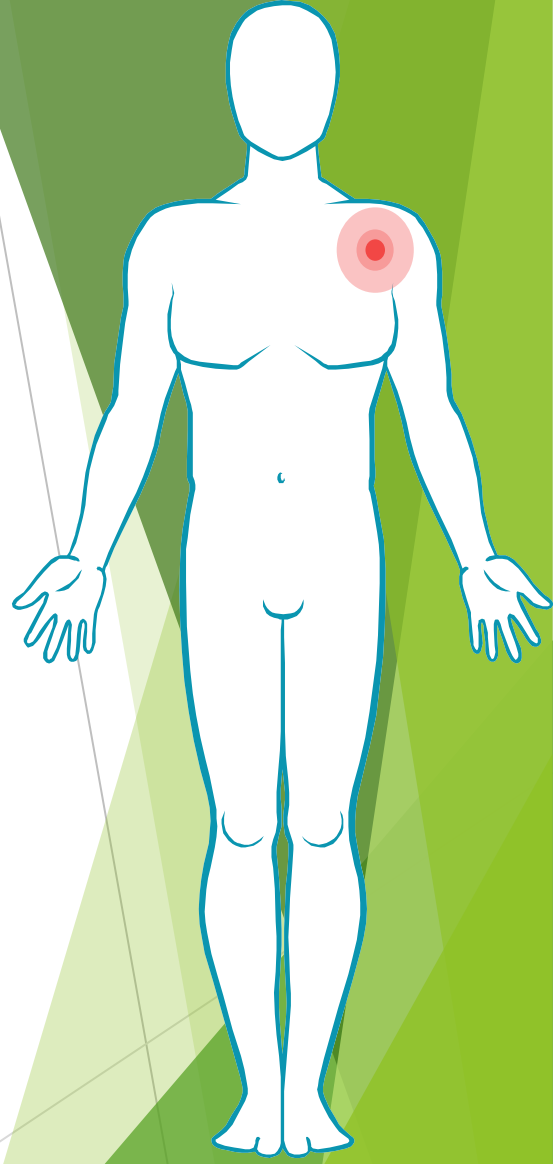
Reference Video

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


Thank
You

DNA Sequencing

Lecture – 4



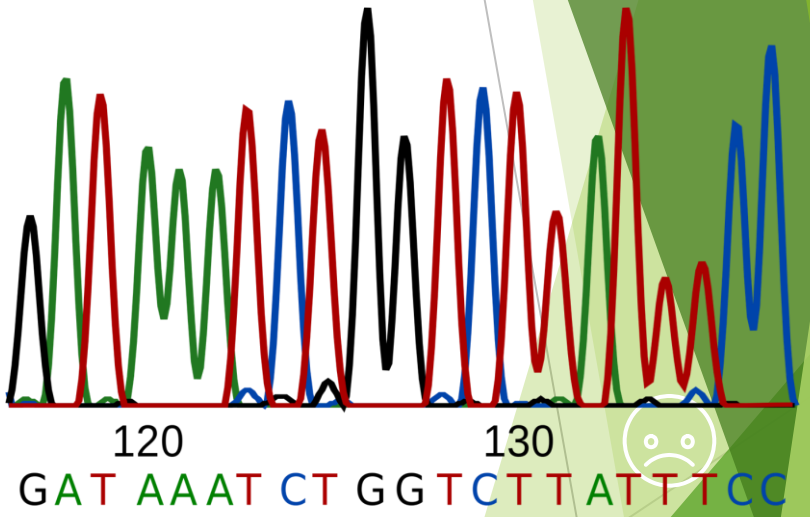
CONTENTS

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.

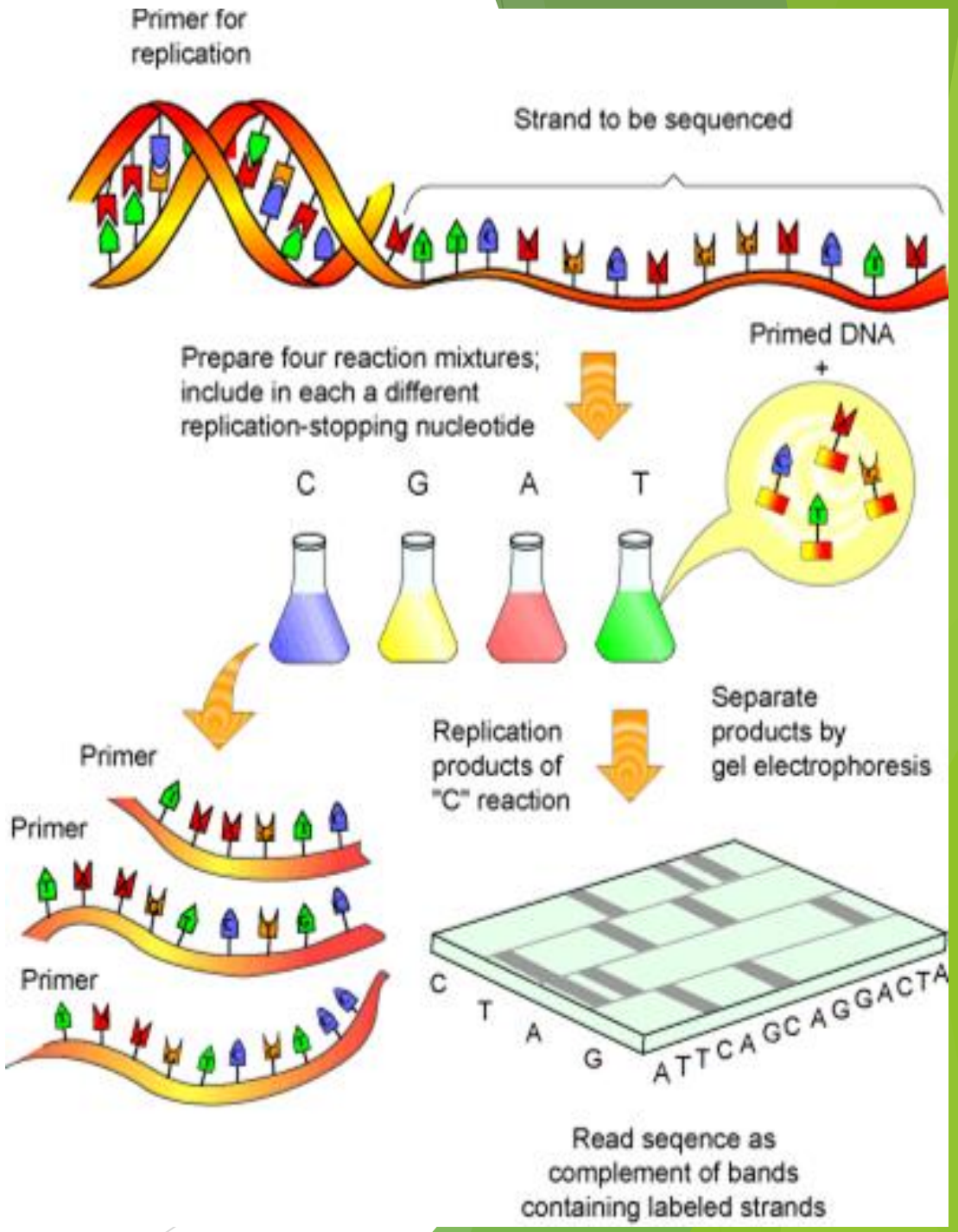


2. First Generation Sequencing

Predominant method for sequencing for decades

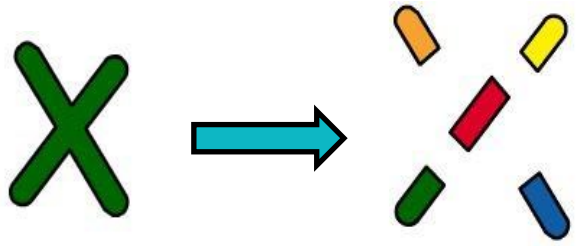
Sanger Method

- 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

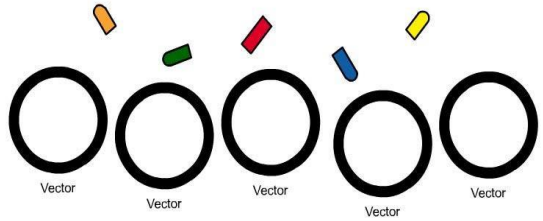


Step 1 - DNA Preperation

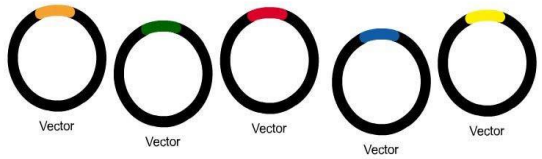
1



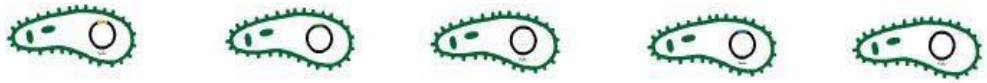
2



3



4



- 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

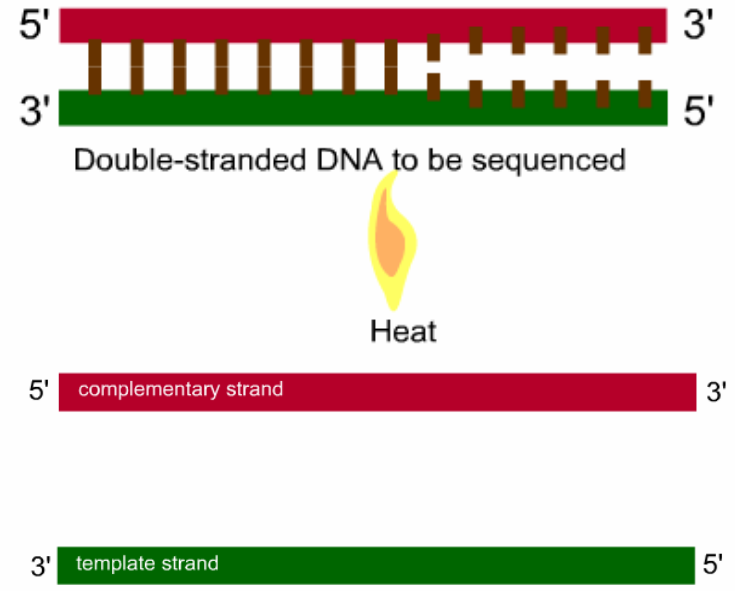
- Strand Separation
- - Heat DNA in 96° C (denaturation)

- 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.

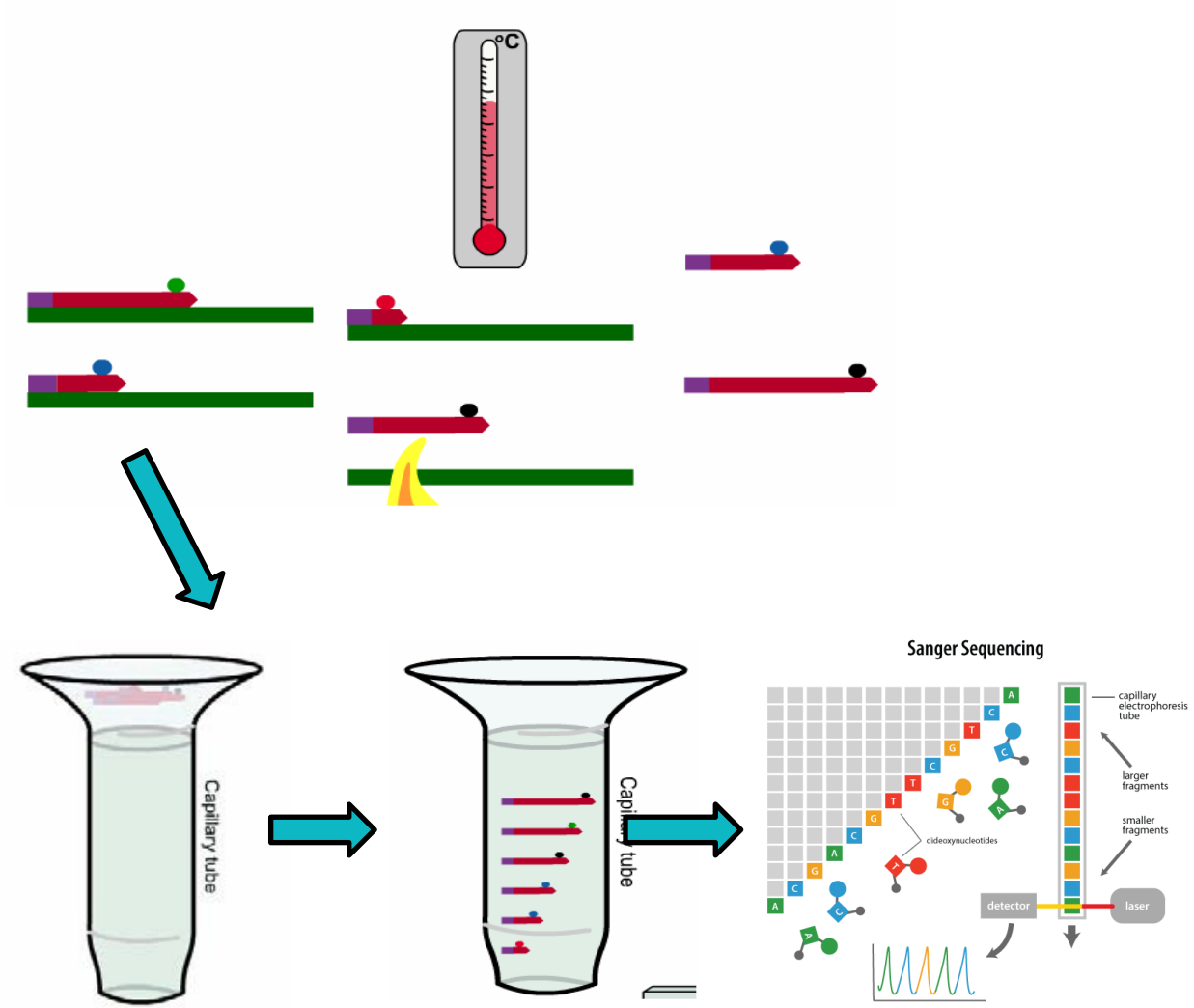


2.

3.

4.

Step 3 – Electrophoresis in Capillary



- Sort the newly synthesized DNA strands by length 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)

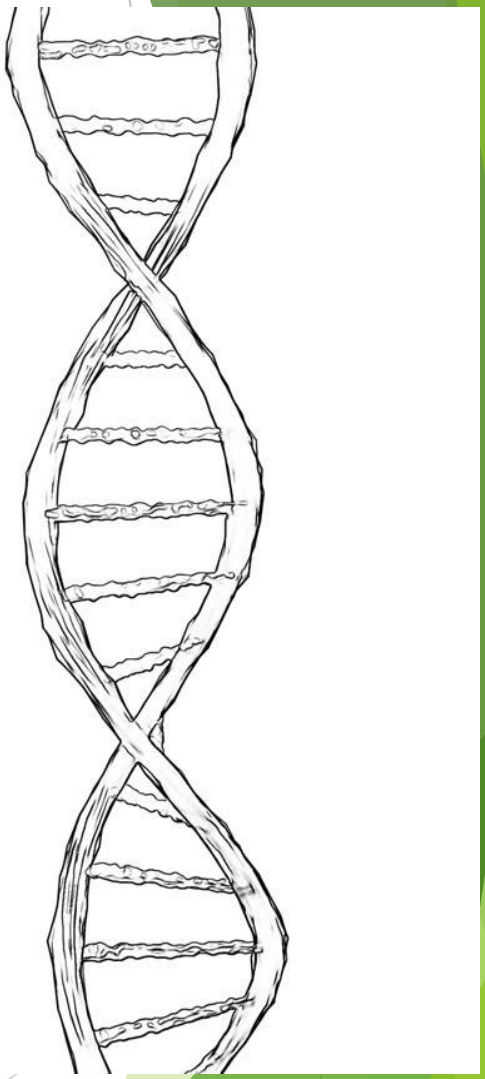
Oxford Nanopore

- 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)

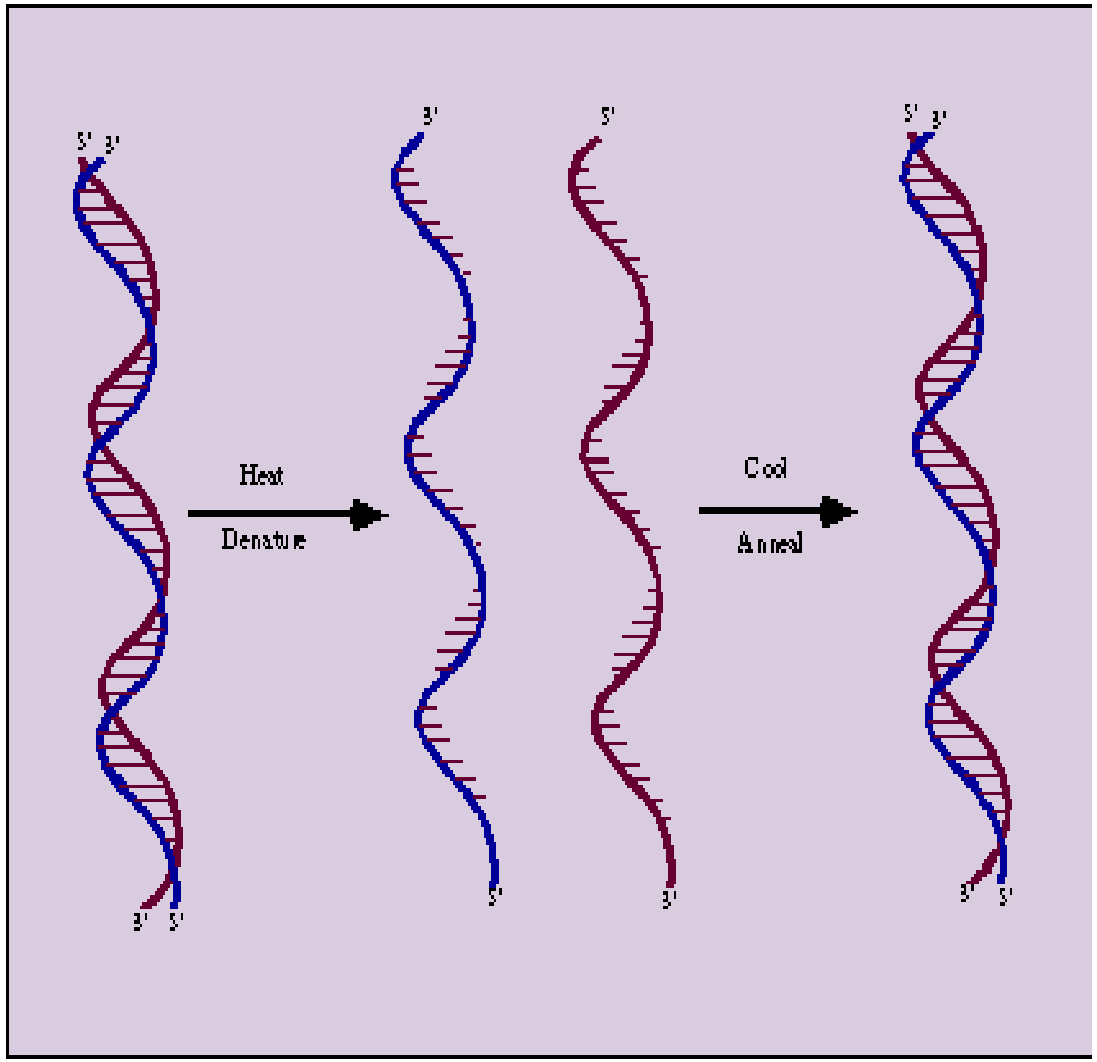
Some comparisons, terms etc.

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

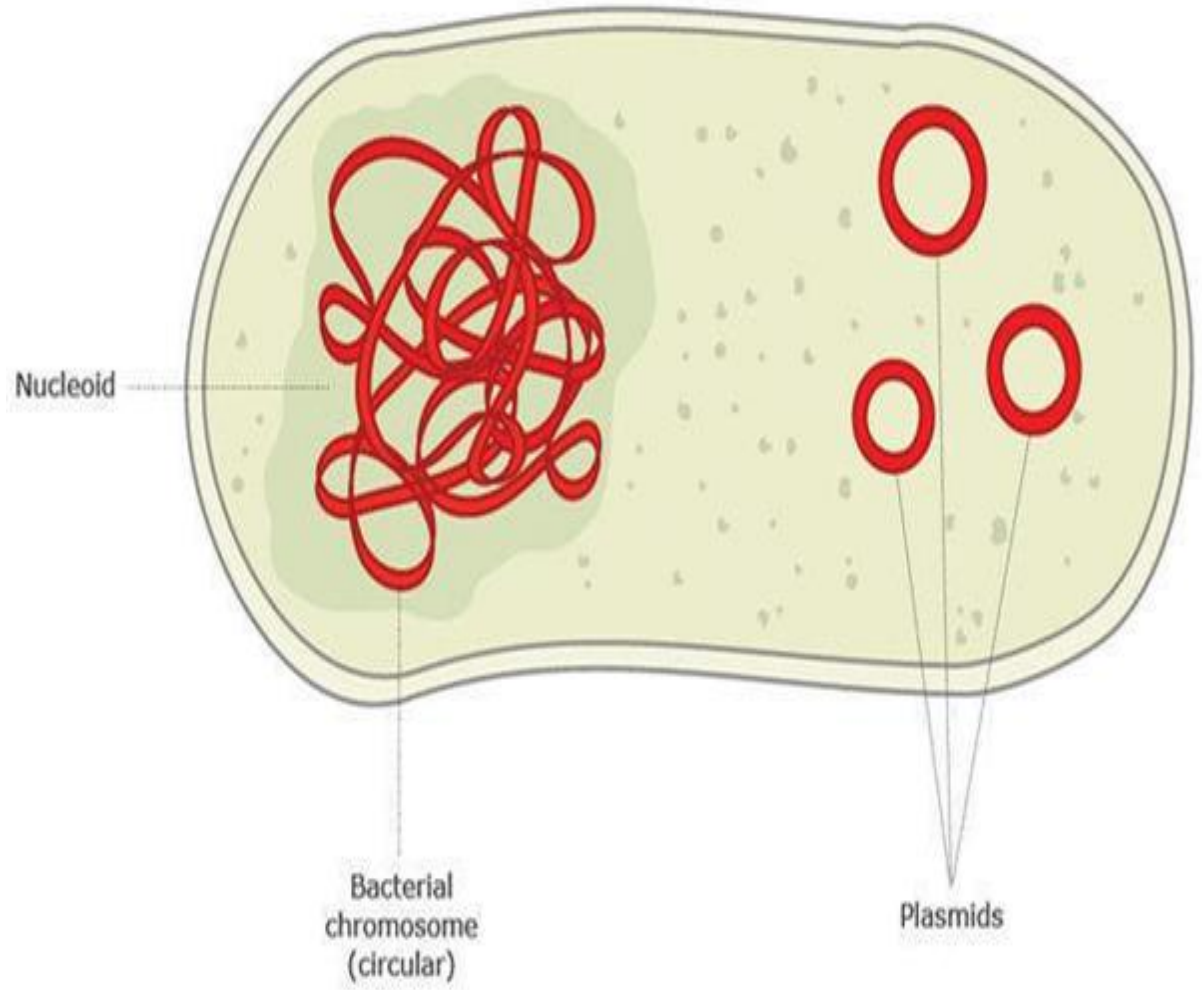
- Small, circular piece of DNA often found in bacteria.
- Sizes of 2.5-20 kb
- Plasmid using method -

* Isolate them in large quantities

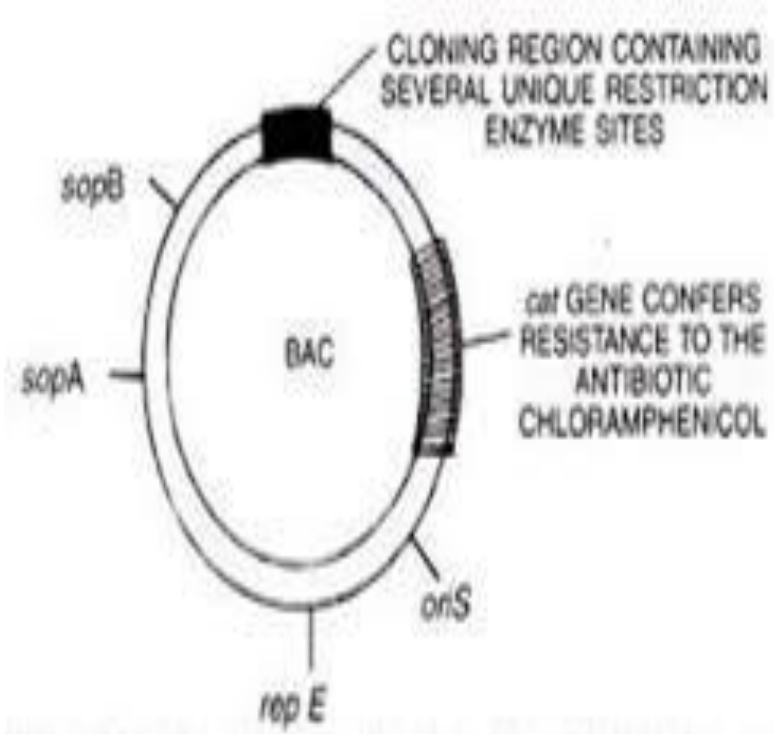
* Cut and splice them, adding whatever DNA needed

* 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



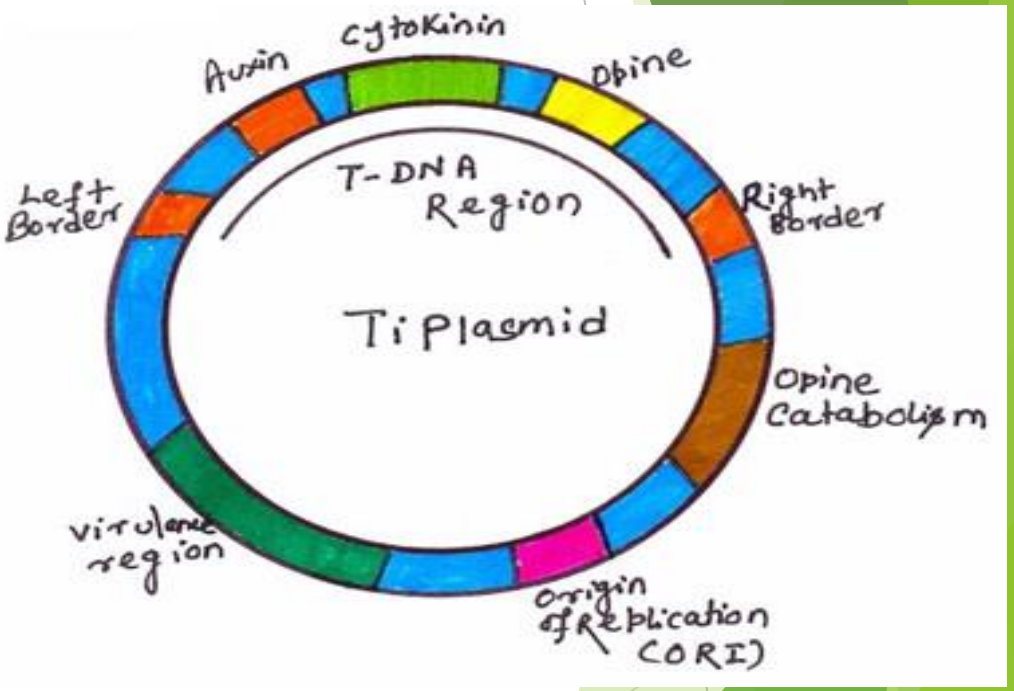
Bacterial Artificial Chromosome (BAC)



- **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 wpm

99.9%

Human DNA is identical, 0.01% creates human diversity

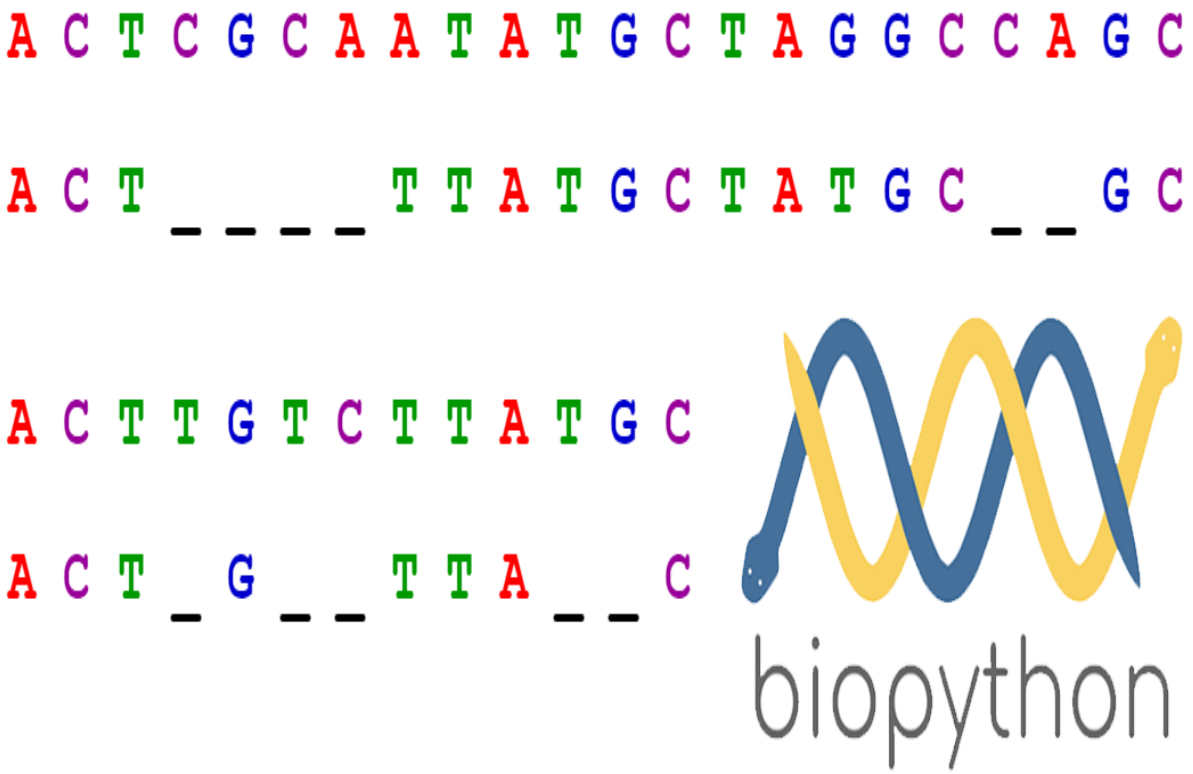
TO BE CONTINUED

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
```

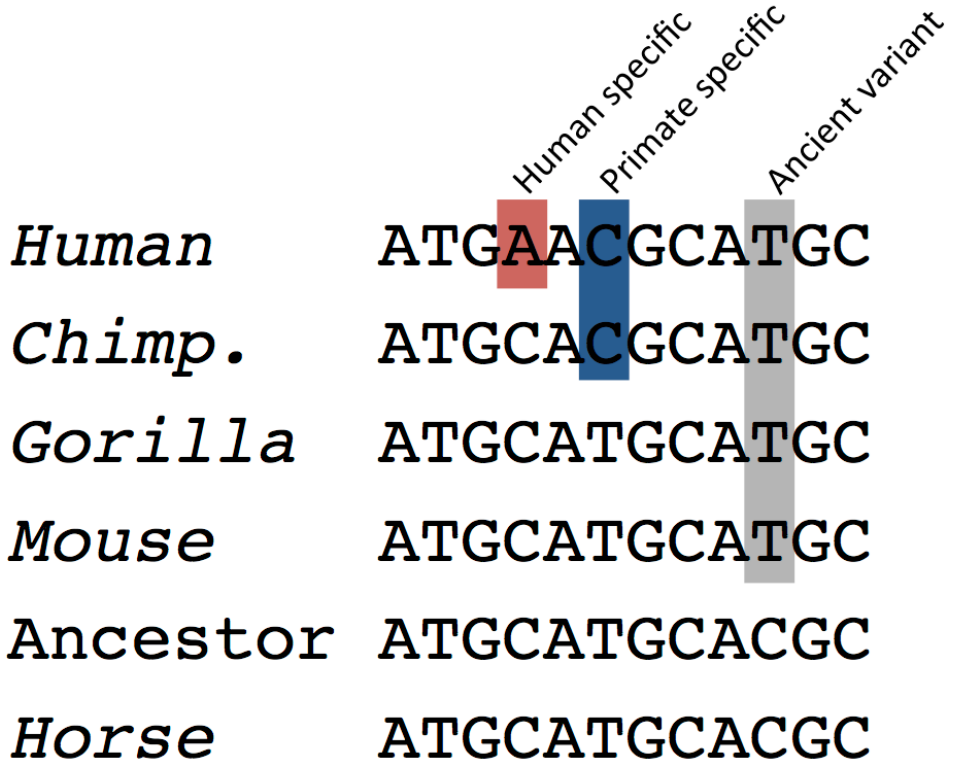
Pairwise Sequence Alignment

- 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 or ← 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

Sequence 1: TTGT

Sequence 2: ATTTGCT

Match Score	Mismatch Score	Gap Score
1	-1	-2

Compute Optimal Alignment | Clear Path | Custom Path

```
A T T T G C T
- T T - G - T
Score = -2
```

		A	T	T	T	G	C	T
	0	-2	-4	-6	-8	-10	-12	-14
T	-2	-1	-1	-3	-5	-7	-9	-11
T	-4	-3	0	0	-2	-4	-6	-8
G	-6	-5	-2	-1	-1	-1	-3	-5
T	-8	-7	-4	-1	0	-2	-2	-2

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
C_2	0	0	0	0	0	4	1	0
G_3	0	2	0	0	0	1	6	3
T_4	0	0	0	2	2	0	3	8

Score: 8

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

TCGT
****
TCGT

Watch:

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


Thank
You