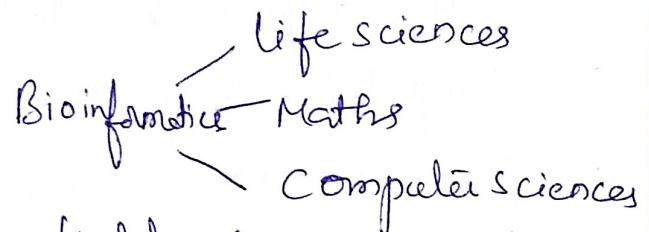


Bioinformatics

Introduction

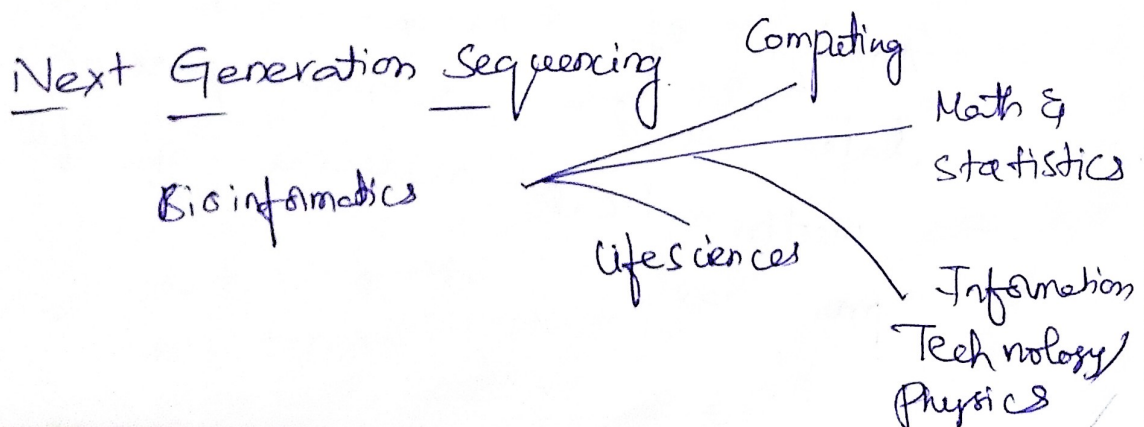
- ① Nucleic acids — DNA & RNA, Heredity
- ② proteins — Catalayst, transporters, structural protein - ~~Strength~~ strength, regulates, metabolism
- ③ lipids — fats, oils, cell signaling
- ④ carbohydrates — Human diet, Carbon sources

→ Databases, Implementation in Computer Sciences



→ Bioinformatics is a field of sciences in which biology, computer science, and information technology merge into single discipline to analyze biological information using computers & statistical techniques.

→ Term bioinformatics was coined by Paulien Hogeweg in 1979 for the study of informatics in biological systems



→ Biological expt can generate large amounts of data (macromolecular sequences, structures, expression.

→ Bioinformatics is about acquiring, managing, analyzing and understanding those data.

Major aspects

- + well organized databases
- Computationally derived hypothesis
- Web servers (tools / online applications)
- Virtual Screening of Compounds for drug development
- Big data (Next Generation Sequence) analysis.

Scope

- Genome and sequence analysis
- From sequence to 3D structural prediction
- Analysis of Genome-wide biomedical data and functional genomics.
- Mathematical modeling of life processes.

- Data base building and management

History:

- In 1978 - Paulien Hogeweg & Ben Hesper coined the term Bioinformatics

- 1987 - Human Genome Project
- 3 billion characters in the DNA
- Sequence - Human genome

- Protein databases - 1st prepared then nucleotide databases

translating into biologically meaningful information
↓
using computer
↓
Bioinformatics

- 1959 - VM Ingram
Compare sickle cell
haemoglobin & normal haemoglobin demonstrates

- 1965 - O. Dayhoff - protein sequences
prepared - Atlas of

→ O. Dayhoff - Prepared Computer methods

- ① In comparing protein sequences
- ② detecting distantly related sequences & duplication within sequences
- ③ deducing the evolutionary histories from alignment - protein sequences

Protein
sequence &
structure

↓
1st published

National Biomedical Research
Foundation

- D. J. Lipman - Director of

National Center for Biotechnology - collaborative project produced BLAST

(Basic Local Alignment Search Tool)

- 1980 - DNA Sequence - EMBL

↓
European Molecular Biology Laboratory

- 1984 - NBRF - Protein Information Resource (PIR)

↓
National ~~Institute~~ Biomedical Research Foundation.

- 1988 - NIH (National Institute of Health)

↓
USA developed NCB (National Center for Biotechnology Information)
↓
As ~~the~~ division of National Library of Medicine (NLM)

↓
Information System of molecular biology

- 1988 - DDBJ, EMBL, Gene Bank, of ~~the~~ International Nucleotide Sequence Database Collaboration had a meeting and agreed to use a common format

- All the databases have collaboration with each other. They regularly exchange their data. Now sequence data are accumulating day by day.

Definition of Bioinformatics

- Fred; Tekala - The mathematical, statistical and computing methods that aim to solve biological problem using DNA & amino acid sequences & related information.

includes

- It ~~includes~~ Computational biology (population modeling & numerical simulations)

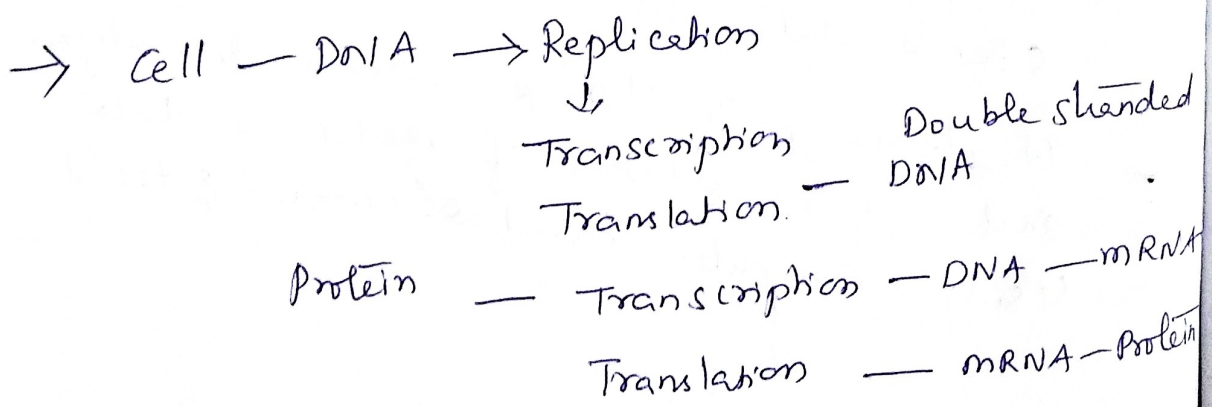
- Bioinformatics is a branch of science that integrates computer science, maths & statistics, chemistry & engineering for analysis, exploration, integration & exploitation of biological science data in Research & Development.

- It deals with storage, retrieval, analysis and interpretation of biological data using computer based software & tools.

objectives

- Development of new algorithms & statistics
 - assessing relationships among large sets of biological data
- Application : Tools for analysis & interpretation of various biological data
- Development of database : storage, access & management of the large body of various biological information.
- Approaches to improving drug discovery ~~and functions in these processes which~~

Central Dogma of molecular Biology



- Protein — sequence (chains of amino acids)
eg:- Lec — Meth — Arg — Try

- Nucleotide sequence

① DNA & RNA

→ Nitrogen bases — A [Adenine]
G [Guanine]
C [Cytosine]
T [Thymine]

- Base pairing — $A \equiv T$
 $G \equiv C$

→ RNA — Instead of T — replaced by Uracil.
Types of RNA — mRNA (messenger RNA)
tRNA (Transfer RNA)
rRNA (ribosomal RNA).

→ Genes — Segment of DNA representing nucleotides.

→ Genome — Total genetic material of an organism.

eg:- Human genome — 3 billion nucleotides

→ Chromatin — DNA + protein — chromatin
(Histone protein)
— Structural unit of chromatin → Nucleosome
— Histone protein & Non-Histone protein
— Nucleosome — solenoids
↓
100,000 bp of DNA.

Applications of Bioinformatics

— Molecular medicine

- ① more drug targets
- ② Personalized medicine
- ③ Preventative medicine
- ④ Gene therapy

— Drug Targets

— Information Search & Retrieval

— Genome Related Applications

— Sequence Comparison

— Linkage Analysis

— Phylogenetic Analysis

— Genomics

— Microarrays

— Sequence Assembly

— Genome Annotation

— Proteomics

— Pharmacogenomics

— Drug-Discovery - Computer-aided Drug Design

— Systems Biology

Applications of Bioinformatics are studied mainly two levels

- ① Basic level, — Bioinformatics is used to organize biological data help researchers access information, add new information arising

out of experiments & modify existing information.

→ 3 datasets $\begin{cases} \text{genome sequences} \\ \text{macromolecular structures} \\ \text{functional genomics expt.} \end{cases}$

→ It is applied in phylogenetic tree analysis, metabolic pathways, medical information, text of scientific papers.

eg:- Protein Data Bank (<http://www.rcsb.org/pdb/>)

(Fast ~~Adaptive~~ Adaptive shrinkage Threshold Algorithm.)
② - Fasta (<http://www.ebi.ac.uk/Fasta3>)

↳ Provide sequence similarity & homology searching against nucleotide & protein databases.

- BLAST (<http://www.ncbi.nlm.nih.gov/BLAST>)
(Basic Local Alignment Search Tool)

- useful to identify distant sequence relationships, statistical interpretation. - Protein & DNA - uses a heuristic algorithm

global alignment

↓
sequences
- Relationships

③ Bioinformatics applications is to use these tools to analyze the data and interpret the result in a Biologically meaningful manner.

① Information Search & Retrieval

— eg:- Pub Med (<http://www.ncbi.nlm.nih.gov/pubmed/>)

↓
National Library of Medicine

↓
provide access to
over 12 million
Medline citations
back to 1960's
+ life science
journals.

eTBLAST (<http://chaos.swmed.edu/e+blast/index.shtml>)

↓
① used to compare query set
sentences with database.

② identify the text database
that is most similar to
the query.

Genetics Related Applications

— 3 types of computational problems in
genetics:

① Analysis of single sequence to assess
similarity with known genes

② Identification of typical features
such as binding sites & derive
evolutionary relationships through

phylogenetic trees.

- ③ Complete genome analysis to identify members of gene families, determination the chromosomal location of gene etc

② Sequence Comparison.

- Sequence comparison ~~or~~ similarity search

eg:- Similarity search - ~~the~~ BLAST & FASTA.

- pairwise comparison of sequences

③ Linkage Analysis

- It identify - chromosomal location of genes, disease identification.

eg:- <http://linkage-rockefeller.edu/>
↓
site for search Linkage Analysis.

④ Phylogenetic Analysis

- It identify evolutionary information to phylogenetic trees.

- Information about phylogeny & Biodiversity

eg:- ① (<http://evolution.genetics.washington.edu/phylip.html>)

② (<http://tolweb.org/tree/phylogeny.html>)

⑤ Sequence Assembly

- Human genome project - popularized sequence - Assembly tools

- used to efficiently & accurately assemble large numbers of individual sequence reads.

eg:-1. <http://www-phrap.org/>

Used to assembly & finishing of shotgun sequencing information.

⑦ Microarrays

⑥ Genomics

→ Term coined by Thomas Roderick in 1986.
- It refers to mapping, sequencing and analysis of genomes.

Structural genomics :- Techniques & tools

1. Linkage analysis
2. Molecular cytogenetics
3. Physical mapping
4. EST sequencing (Expressed sequence tags)
5. Genome Sequencing
6. Genome organization

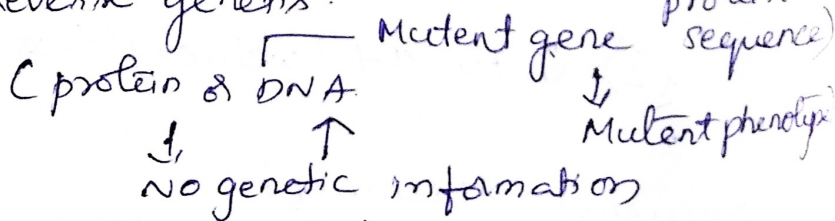
Functional genomics :- To study gene function

Tools :-

① Gene expression

② Forward genetics (gene → DNA

③ Reverse genetics.



④ Comparative genomics

⑤ proteomics

⑥ Metabolomics

Unit-2

Introduction to Biological data bases

→ Biological Database.

Bio-life, Database - collection of data & arranged & managed

→ Biological Database is the a collection of biological information that is organized so that it can be easily accessed, managed & updated.

→ Data :-

① DNA sequence of gene or full genome.

② sequence of protein and 3-D structure of protein, nucleic acids and protein-nucleic acid complex, transcriptomic

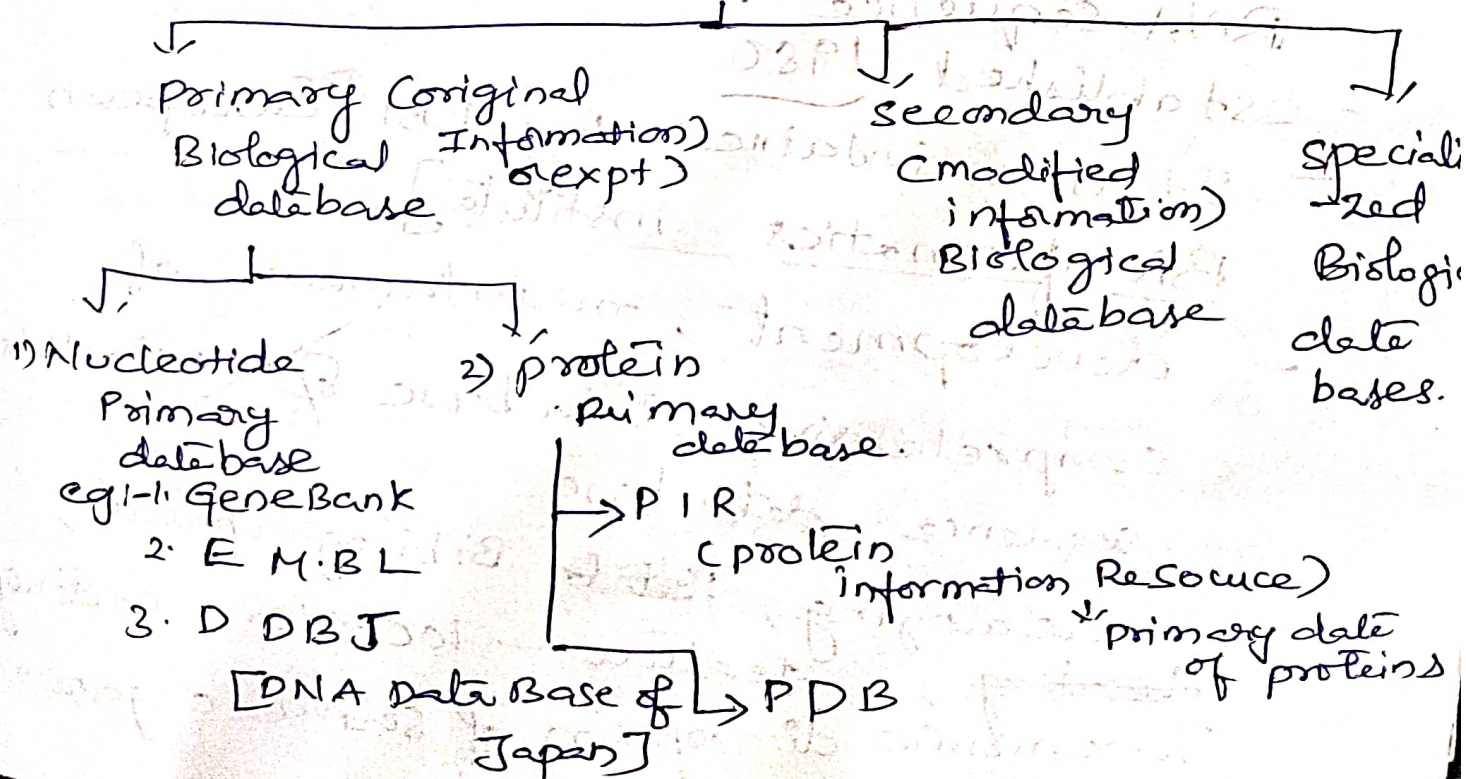
- There are 15,52 database - publically accessible

- Genomics, proteins.

- Bibliographic databases - Literature

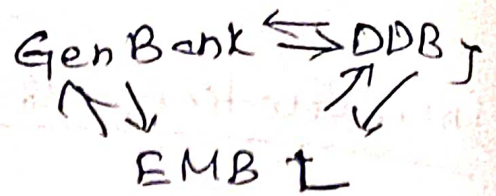
- Taxonomic databases - classification

Biological data bases



Gene Bank

- Nucleotide data
- Primary data
- It has Flat file structure.



↓
It access text file

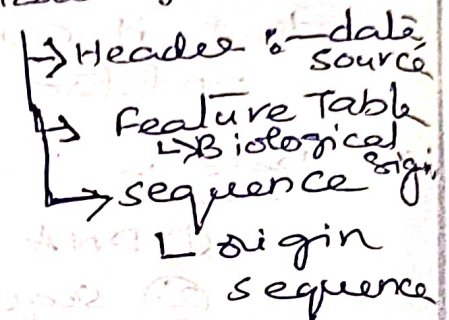
- NCBI, MD, USA - data managed.

Gene Bank Records

eg:- Flat format

→ DDBJ is located at

- National Institute of Genetics (NIG) in Japan



- DDBJ began data bank activities in 1986 at NIG and remains the only nucleotide sequence data bank in Asia.

EMBL

- A comprehensive database of DNA and RNA sequence.

- established 1980

- It is maintained by EBI [European Bioinformatics Institute].

- development, maintenance of comprehensive data base of nucleotide sequence data base
- cataloging, ~~feature~~ Bibliographical, ~~texting~~ feature - location, coding Taxonomic description of source organism.

region, Biological significant site.

swissprot

- This is a curated protein sequence database that offers a high level of integration with other databases and also has very low level of redundancy.

→ It strives to provide protein sequence with a high level of annotation (for instance, the description of protein function, domain structure and post translation modification etc).

TrEMBL - Translation (Tr)

→ TrEMBL is a computer-annotated supplement of swissprot that contains all translations of EMBL nucleotide sequence entries, which is not yet integrated in swissprot.

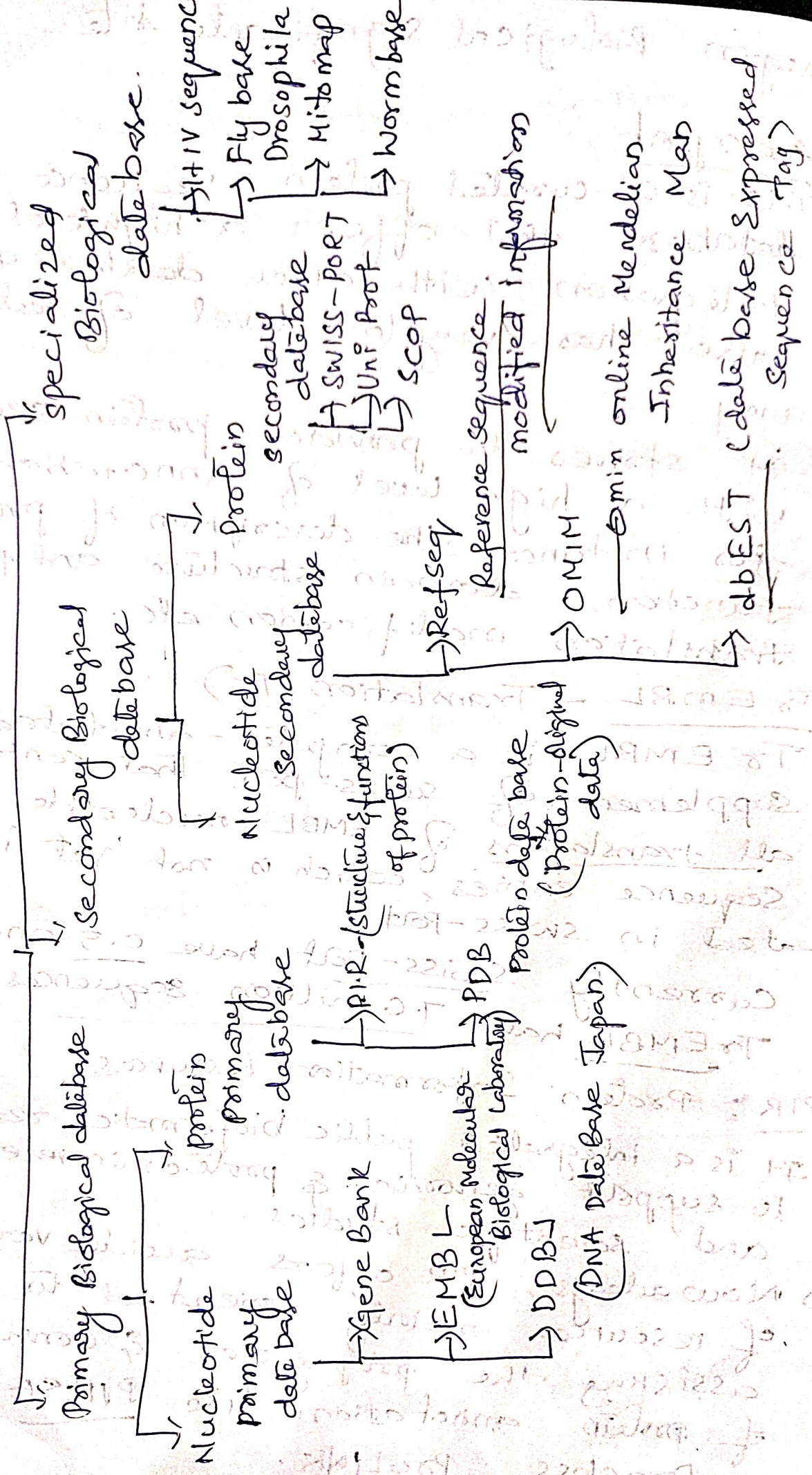
→ Currently swissprot have 0.5 and TrEMBL have 7.6 million sequences.

PIR: Protein Information Resources

→ It is an integrated public bioinformatics resource to support genomic & proteomic research and scientific studies.

→ Nowadays, PIR offers a wide variety of resources - mainly oriented to assisting the propagation & consistency of protein annotations like PIRSF, ProClass, ProLink.

Biological databases



Structural database :-

- SCOP — structural classification of protein
 - ↳ 3D structure of protein
 - ↳ Classified
- CATH — class, Architecture, Topology, Homology
 - ↳ hierarchical classification of protein
- GEO → Gene Expression Omnibus is a gene expression resource and a gene expression
- GXD → molecular abundance
- MGED → repository for gene expression data browsing query & retrieval.

Gene expression database

GXD

- gene expression data
- community resource for gene expression information

MGED

- Microarray Gene Expression data
- contain microarray data, generated by functional genomics and proteomics. Expt

metabolic → Kyoto Encyclopedia of Genes and Genome
data base → KEGG — pathway database — metabolic
base → LIGAND — Kyoto — chemical Research. pathway
base → Chemical data base.

EcoCyc — E. coli. Enzyme data, repair,
Eg: — Database model organisms drug, Glycan compound

① E. coli → ① E. coli Genome Centre
(Wisconsin University USA)

② Arabidopsis thaliana → ② E. coli Index
(Birmingham. UK)
TAIR

③ Homo sapiens → HGR at NCBI, USA

④ Oryza sativa → RGP (Japan)

⑤ *Drosophila melanogaster* → Fly Base

⑥ *Mus musculus* → Mouse Genome Informatics
(mouse)

⑦ Danio rerio → ZFIN (zebrafish
Information
Network
Oregon, USA.

⑧ *S. cerevisiae*. $\rightarrow$ Yeast Genome Database - Stanford
US A

Secondary data base

① PDB :- Protein Data Bank

① PDB :- Protein Data Bank

- 3D structure - macromolecules - X-ray crystallography
- Expt. data - structures - NMR
- homology model

→ RCSB PDB — Research Collaboration
Structural ~~B~~ data base.
tool — structure, relationship,
function, diseases causes

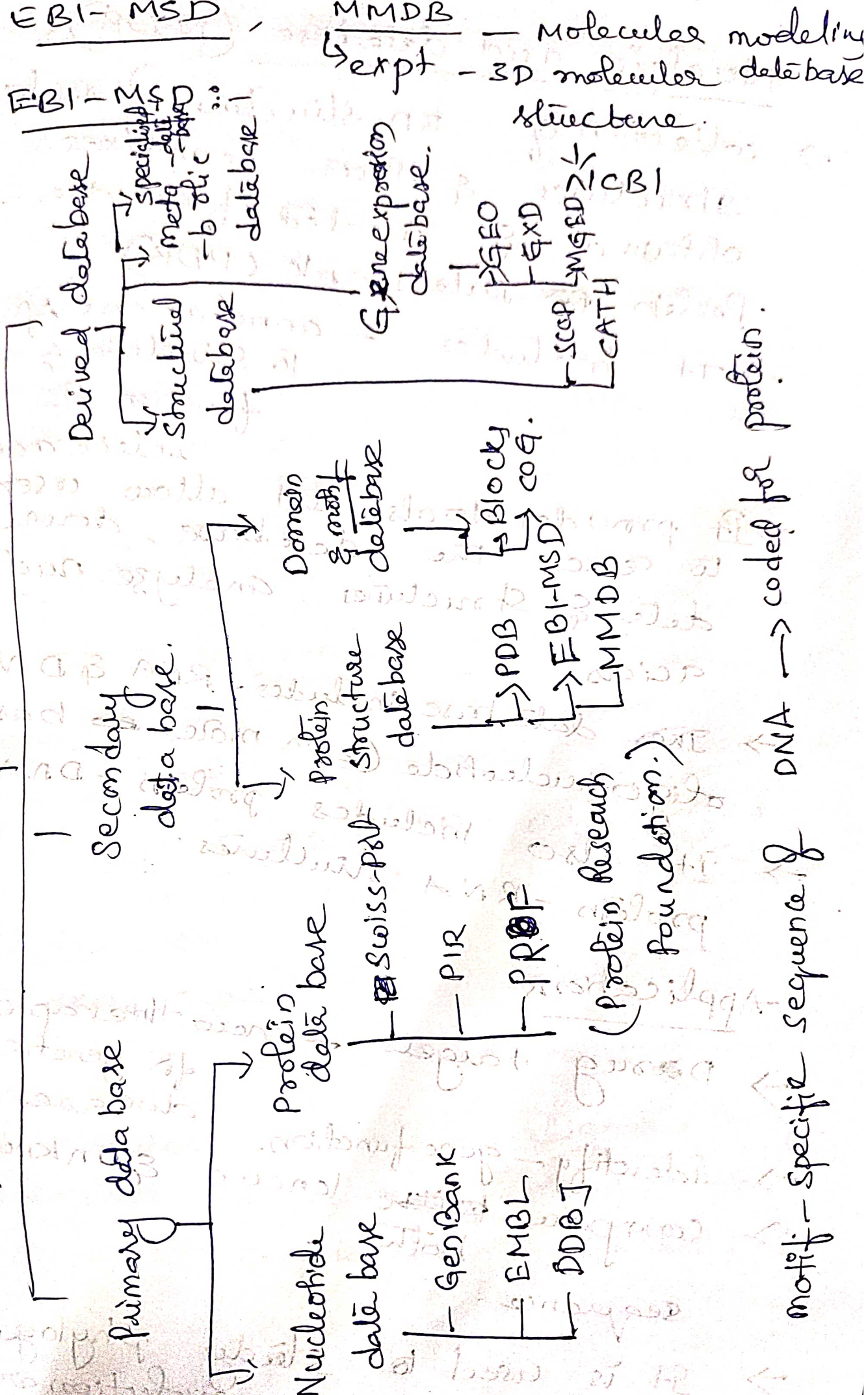
gateways to PDB files.

Swiss-Prot, NCBI, EMBL

protein data Bank.

CATH, DALI, SCOP, PDB
databases, ~~that~~ interpret PDB files

Biological data base



② EBI-MSD

MMDB — Molecular modeling
 Expt — 3D molecular database structure.

② EBI — Specialized meta-database

DNA → coded for protein.

motif — Specific sequence of

Nucleic acid database (NDB)

- collection of 3D structure of nucleic acid structures & their complexes obtained & curated from the Protein Data Bank (PDB)
- It includes - annotations specific to structure & function of nucleic acid.
- It provides tools that allow users to search the database, download data & structures, analyze nucleic acids.
- The database includes RNA & DNA oligonucleotide (2 or more ~~or~~ bases)
- It also includes protein-DNA & protein-RNA structures.

Applications

- Drug targets - new therapies for genetic diseases.
- Identify - gene function.
- Compare ~~both~~ known & unknown sequence.
- It is used to study phylogenetic analysis to analyze evolutionary

- relationships b/w organisms -
- comparing & examining their DNA & RNA sequence.
- It used to identify gene expression.
-

file format - Genebank

① Headers

- locus
- Definition
- Accession
- Version
- Source
- Keywords
- organism
- References
- Authors
- Title
- Journal
- Medline ID : all published sources
- Comment

② Features

③ sequences .

BLOCK :

- It contains multiple alignments of conserved regions in protein families.
- eg:- Google cloud Persistent Disk,

COG - clusters of Orthologous Genes

- 1997

- Identify protein sequence

→ phylogentic classification of the proteins encoded in 21 complete genomes of bacteria, archaee & eukaryotes.

→ 1,187 bacteria and 122 archaea mapped

→ uses :- Comparative evolutionary studies and enables the most specific inferences of tentative gene functions.

→ Uniport :- universal protein database

→ Ref Seq :- Reference ^{sequence} modified information

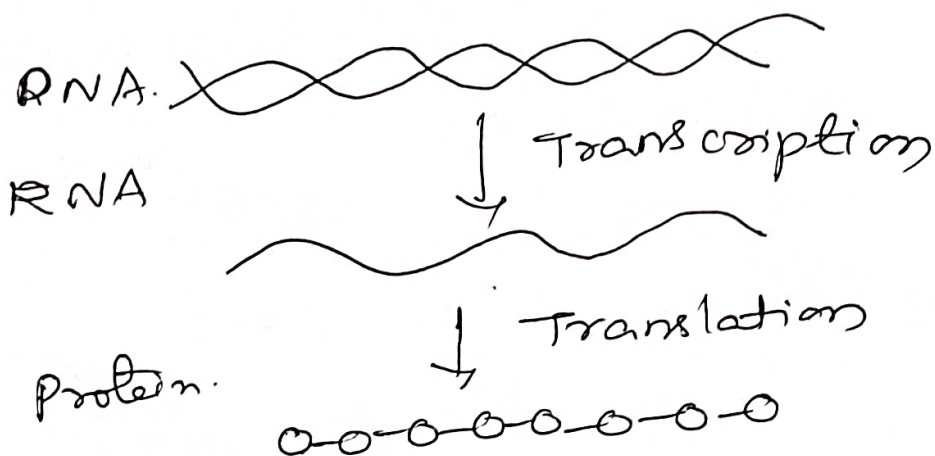
→ mito map :- Mitochondrial mapping

What is Sequence?

→ An arrangement of 2 or more related things in a successive order

Related things:-

Here there are DNA (base pairs as (A C T G)), RNA (base pairs A C U G & proteins amino acids)



Sequence Alignment

→ The alignment of biological sequence is probably the most important and most accomplished in the field of bioinformatics.

→ Their ultimate goal is to determine the similarity b/w different sequence.

A	G	C	T	G
A	G	C	(A)	G
A	G	C	T	G

Sequence Similarity Vs Sequence Identity

① For nucleotide sequence DNA & RNA
Both words have same meaning

- eg:-



Two DNA sequence can share a high degree of sequence identity & similar - meaning same

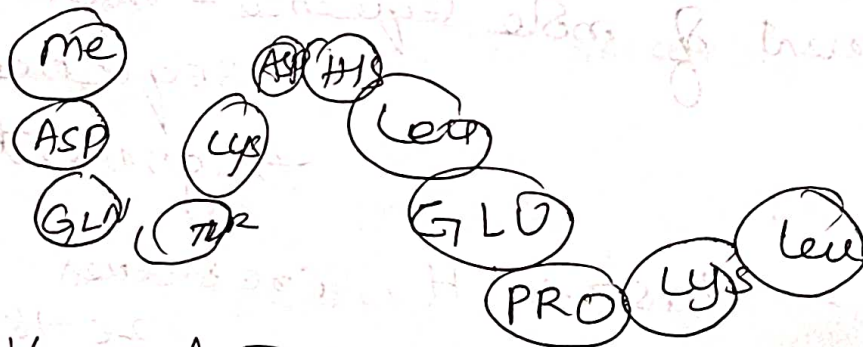
② for protein sequences:-

Both words have different meanings

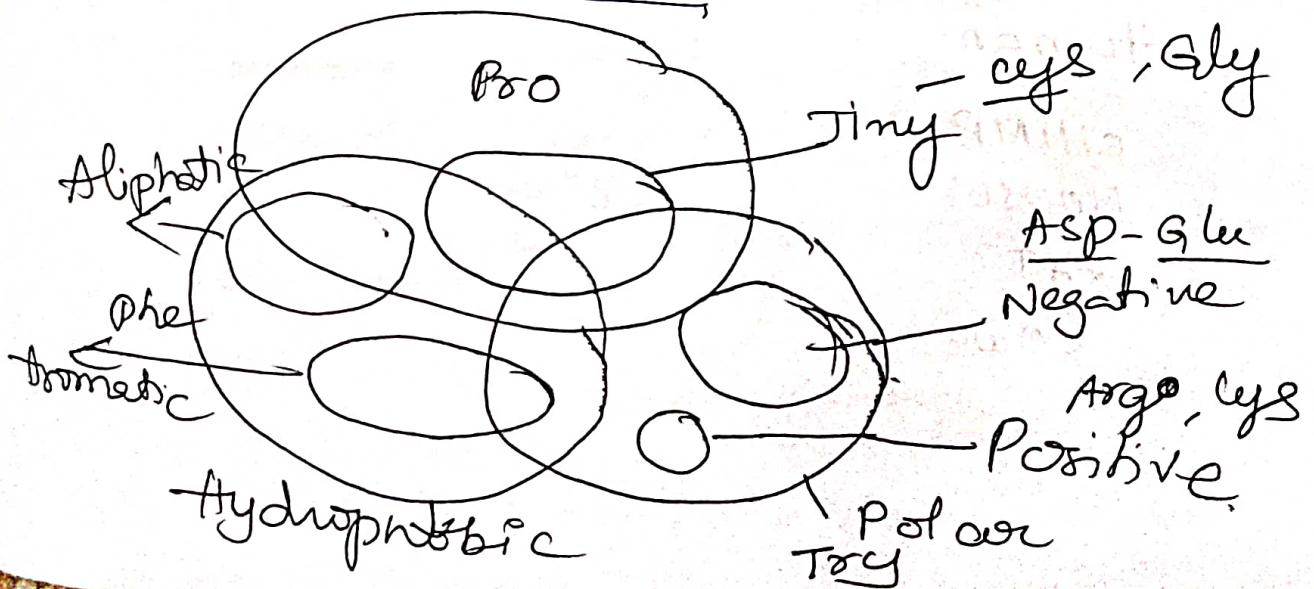
Identity :- % of exact matches b/w 2 aligned sequences

Similarity :- % of aligned residues that share similar characteristics

eg:-



Similar characteristics



eg: Identity - ?

If sequence $A=B$ & $B=C$?

A be equal to C. (In terms of Identity distance means)

Strand A $\rightarrow$ A A G G C T T

Strand B $\rightarrow$ A A G G C

Strand C $\rightarrow$ A A G G C A T

- A, B = 100%, identical. - 5 nucleotides

B, C = 100%, A, C = ~~88%~~ (88% ^{show ident})

- So 100% identity does not mean two sequences are the same. 6 identical nucleotides

How Sequence Alignment
Alignment of 2 sequences — Pair-wise Sequence Alignment.

$\rightarrow$ Alignment of multiple sequence — multiple Sequence Alignment.

eg: - Histone H, (residues 120-180)

Human.

CHIMP

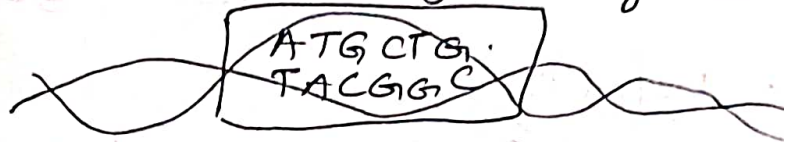
MOUSE

RAT

Cow.

why sequence alignment

- gene finding - identify open reading
frame (ORF) of genome
- length of gene, probability
actually constituting
gene



→ function prediction

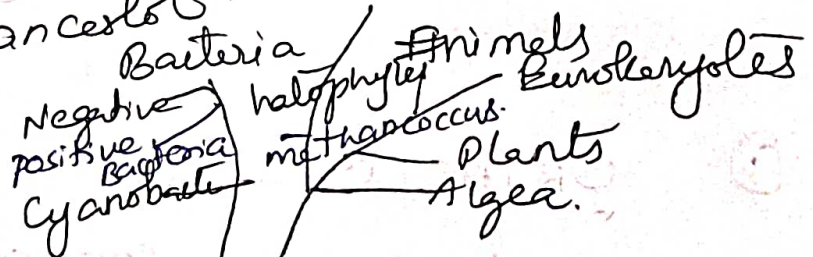
- gene - eg: Protein - functions.
enzyme
eg: Insulin, haemoglobin, RUBiSc,
- actin & myosin [contraction]
- ~~storage~~ ferritin (storage)
- immunoglobulin (protection)
Antibiotics.

→ Assembly, Genome :-

- Base alignment short DNA sequence
obtained by new
- Identification of homology gene

sequencing → quality control → Assembly → Annotation
Homology ↓
Comparator

- Both identity and similarity are used to deduce
homology. Homology, however has a
specific definition - having a common
evolutionary ancestor.



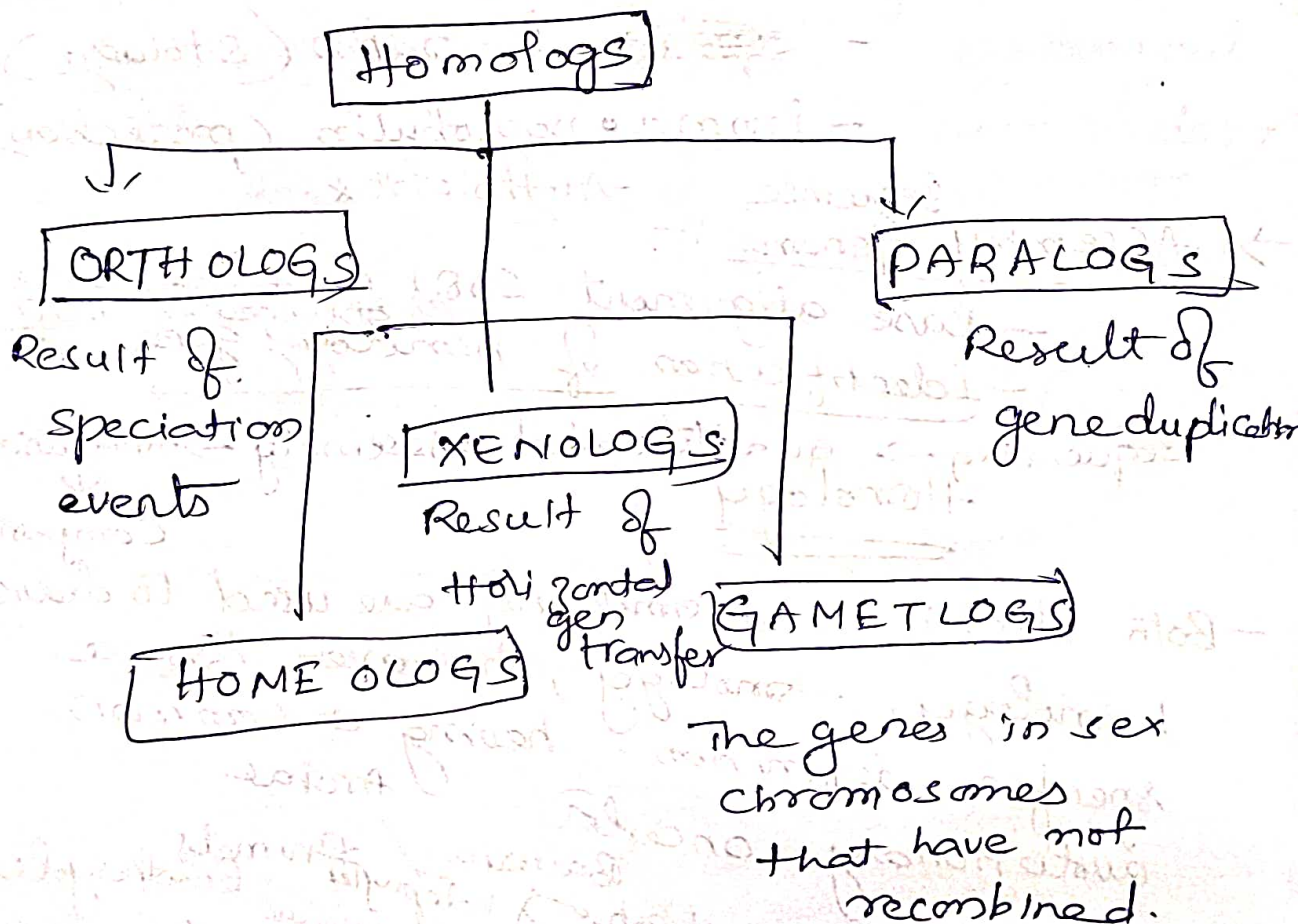
last common
Ancestor

→ ~~Both~~ homology is defined as similarity b/w organisms ~~of~~ of gene which have common ancestor, physiology, identity, relationships.

Homology

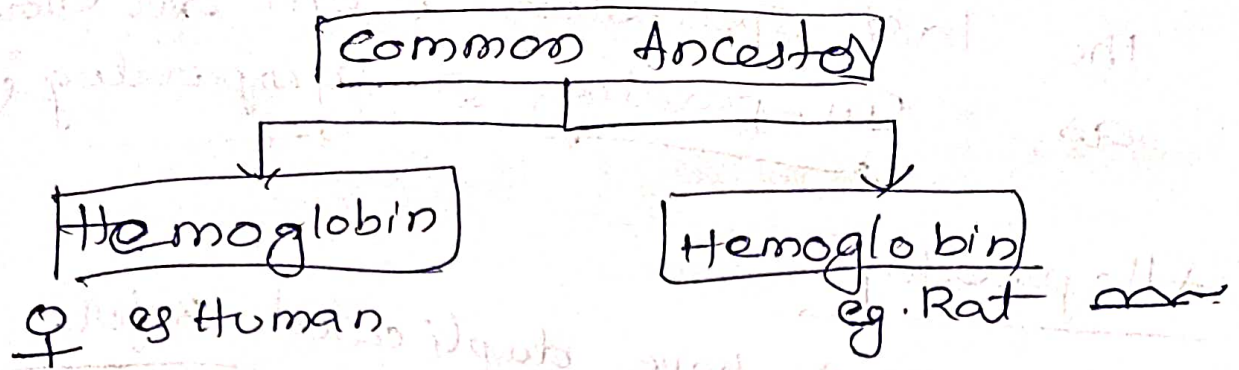
- Definition :-

Homology are two or more sequences that descend from a common Ancestral Sequence. Homologs are results of divergent evolution.



① ORTHOLOGS :-

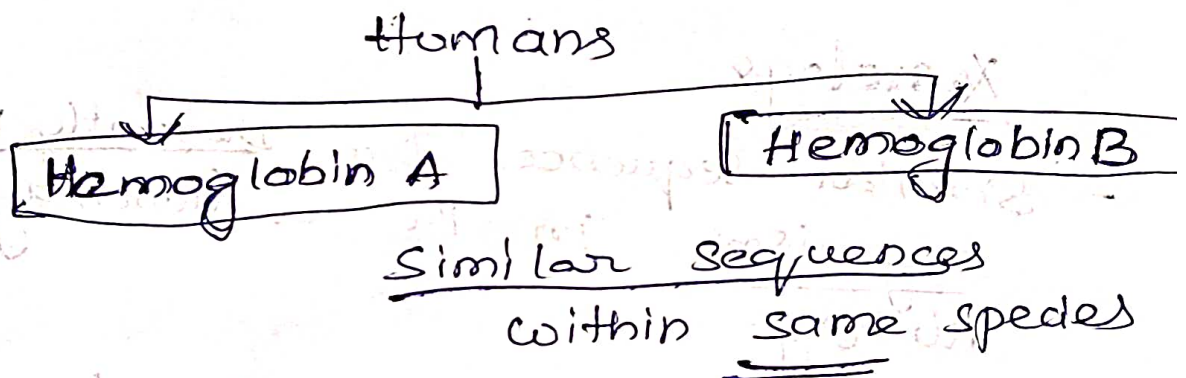
- Orthologous sequences are those sequences which are the results of the speciation process.



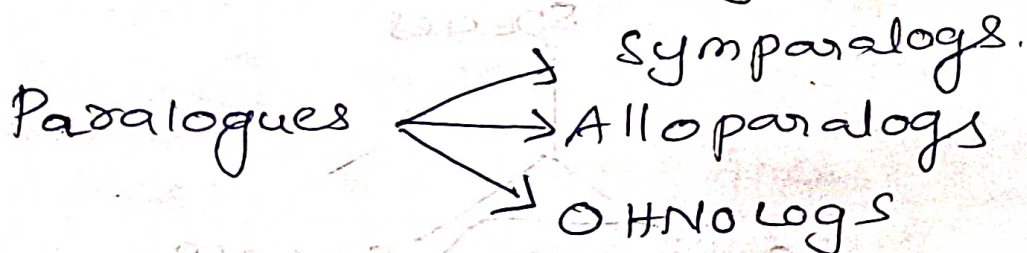
- Similar or identical sequences useful in different species
- construction of phylogenetic trees, information of species, evolution of genomes

PARALOGUES

- Paralogous sequences are homologous sequence that have diverged within one species. i.e. arise during gene duplication process

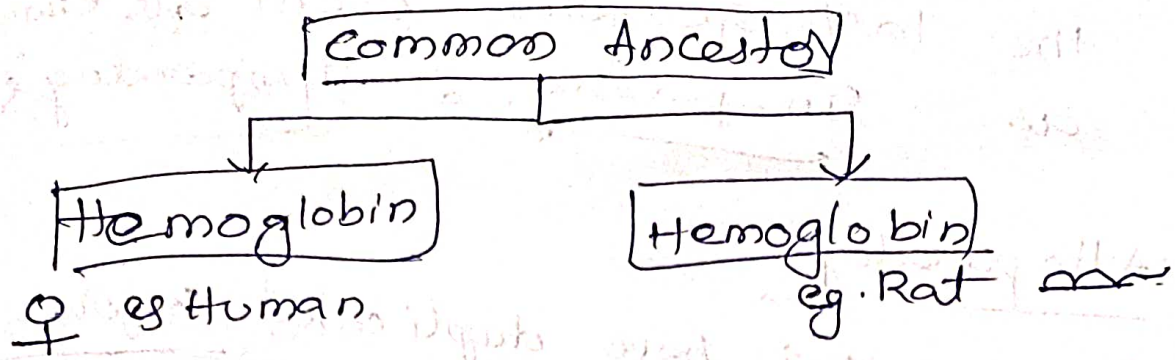


- protein evolution — Imp
- eg: — Hemoglobin A, B, A₂, B, F
- paralogues genes.



Symparalogs

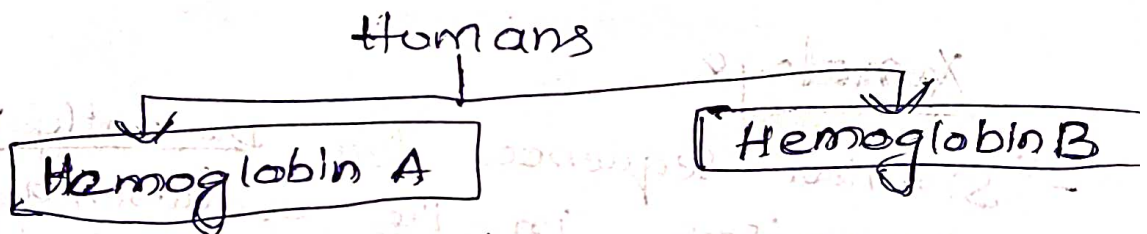
- Genes that have duplicated After



- Similar or identical sequences useful in different species
- construction of phylogenetic trees, information of species, evolution of genomes

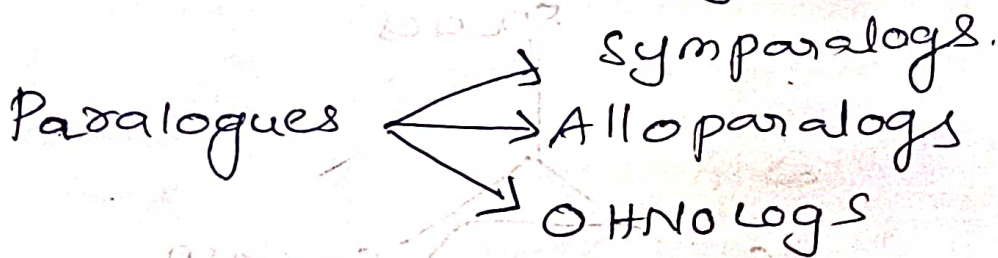
PARALOGUES

- Paralogous sequences are homologous sequence that have diverged within one species. i.e. arise during gene duplication process



Similar sequences
within same species

- protein evolution — Imp
- eg: — Hemoglobin A, B, A₂, B, F
- paralogous genes.



Symparalogs

- Genes that have duplicated After

the last speciation event are known as In-paralogs or symparalogs

2. Alloparalogs

→ Genes that have duplicated Before the last speciation event are known as Out-Paralogs or Alloparalogs.

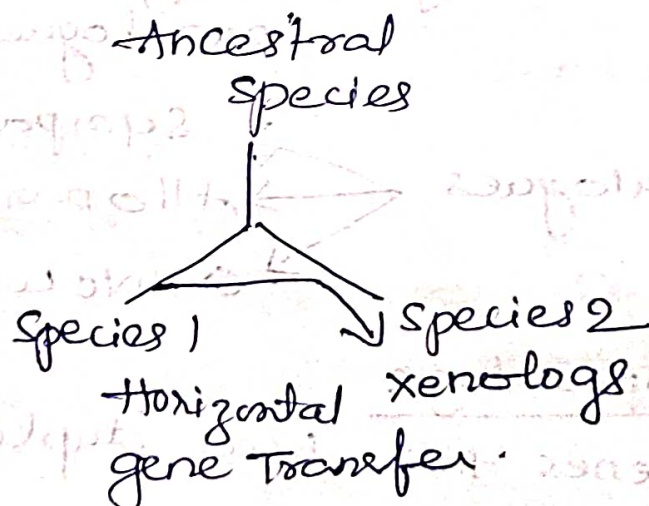
3. Ohnologs

→ these are genes that are a result of a duplication of a whole genome

Xenologs

- similar sequences b/w Distantly related organisms in the evolutionary history.

Horizontal gene Transfer
lateral gene Transfer

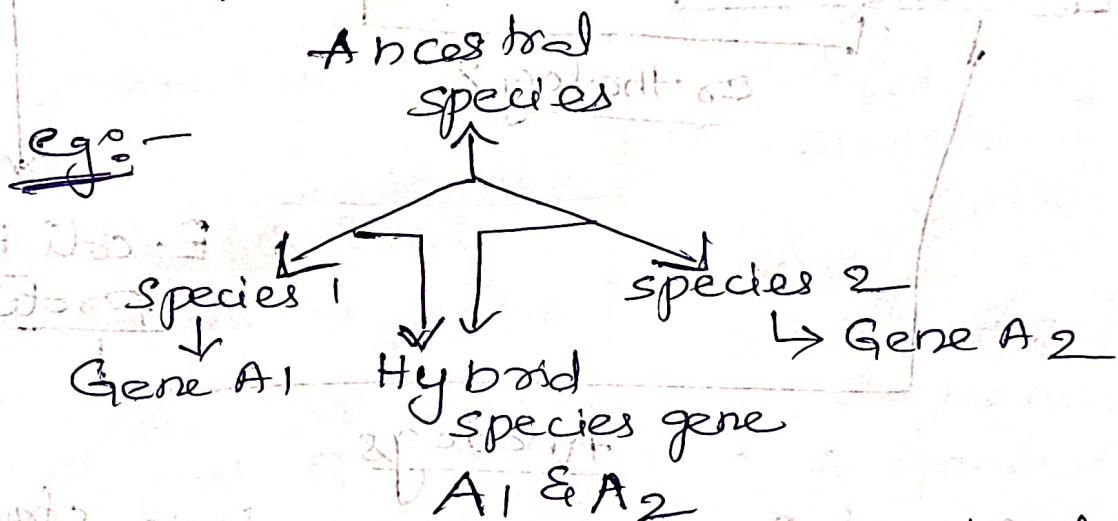


→ Lateral gene transfer → disease pathogens
 ↑
 Non-related organisms
 ↓
 Out breaks

Homeologs

→ ~~Re~~ Homeologs are special cases of xenologs.

→ • The genes which are separated by a speciation event when hybridized together via lateral gene transfer, the gene sequences are known as Homeologs



- Bilateral gene transfer — Hybrid: ↓ Homeologs.

Analogous :-

→ Analogous genes are genes that have identical or similar functions but Don't

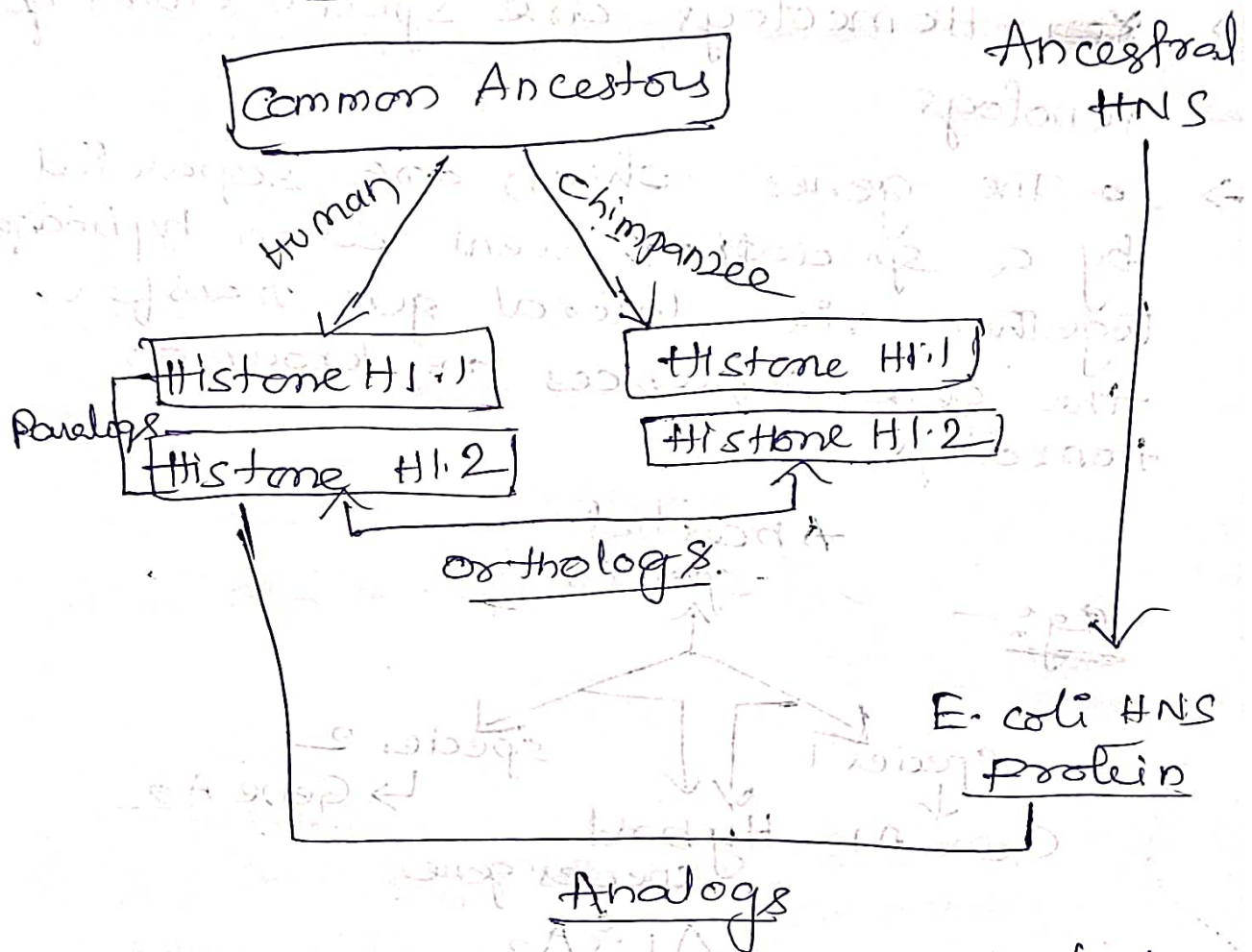
share a common Ancestor sequences.
 convergent Evolution

eg: Squid eye Human eye
 ↑
 fishes

→ Analogs have homologous activity
But heterologous origin.



Homologs vs. Analogs



HNS is Histone-like nucleoid structuring protein.

Types of Sequence Alignment

alignment of two or more sequence.

- ① Pair-wise sequence alignment
- ② Multiple sequence alignment

→ Pair-wise: alignment is used to identify regions of similarity that may indicate functional, structural & evolutionary relationships b/w two biological sequences.

eg:

A	C	G
A	T	-

Match Mismatch Gap

① DNA
Nucleotide Base
sequences

② Protein
Amino acid
sequence

Why alignment the sequences?

- functional Relationship
- Structural Relationship
- Evolutionary Relationship

~~Also~~ Alignment

Global Alignment

- ① Tries to align entire sequence
- ② Align all letters from query and target
- ③ Suitable for closely related sequences
- ④ A general Global Alignment method is Needleman-Wunsch

Local Alignment

- Align similar sequence
- Align substring of target with substring of query
- suitable for more (distantly related) divergent sequences
- + A general local Alignment method is Smith-Waterman

Global Alignment: 1970

Rewards & Penalties

Match = 1

Mismatch = -1

GAP = -2

Both needleman & Smith waterman

→ Dynamic programming

→ steps simple problems

→ steps

Steps

① Initialization $\rightarrow$ matrix filling

④ Alignment ③ Traceback

→ Algorithm of two sequences → Needleman-Wunsch

Seq. 1 - $\frac{A T G C T}{M}$

$$Seq_2 - \frac{AGCT}{N}$$

		A	T	G	C	T
	O	-2	-4	-6	-8	-10
A	-2	1				
G	-4					
C	-6					
T	-8					

十

Match = 1
Mis match = -1
Gap = -2

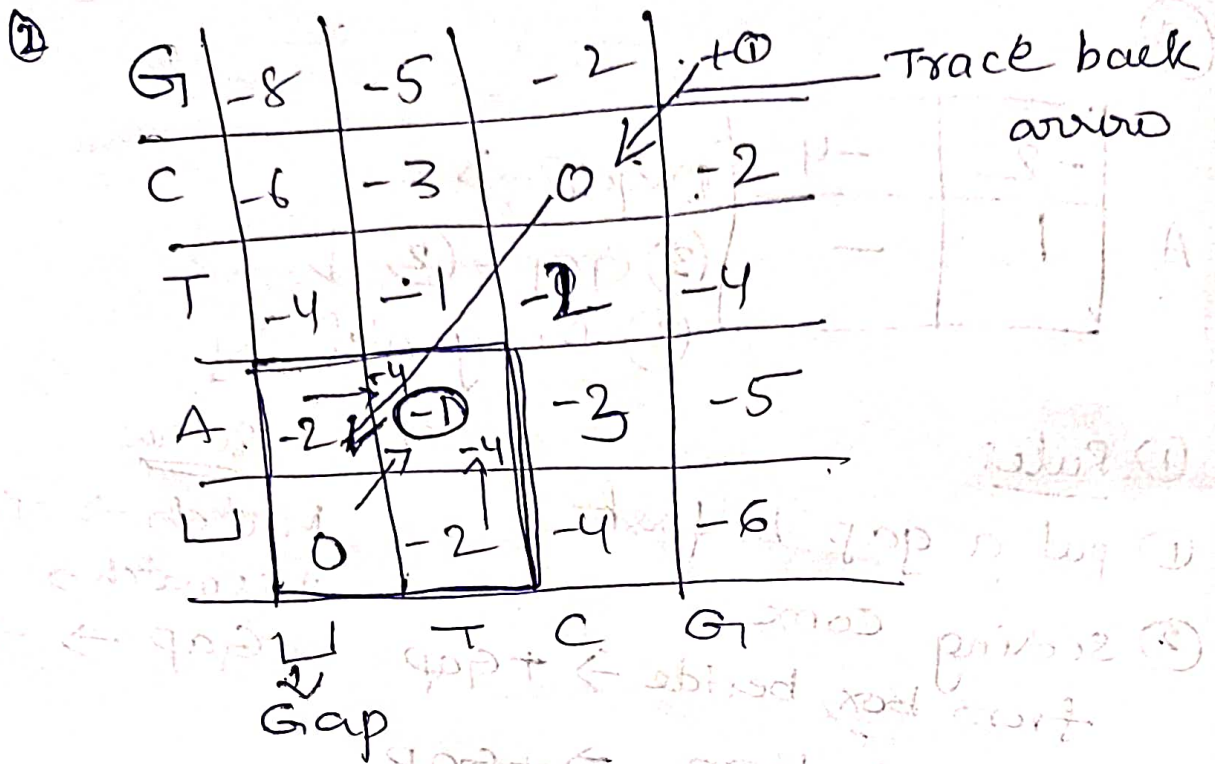
①

$$A = \begin{bmatrix} 0 & -2 \\ -2 & -2 \end{bmatrix}$$

value from left = -4

Value for $u_p = -4$

$\checkmark$ value from diagonal = 1



Rules for Alignment of sequences

- Track back arrow
- Direction of arrow →
 - ↑ → Gap
 - ↖ → diagonal (characters)
- direction of arrow
 - vertical or horizontal put - Gap.
- characters - sequence characters

- ① Needle man - Wunsch algorithm
 - ② Smith - waterman algorithm
- } Dynamic programming

Local sequence Alignment, 1981

- smith waterman

- Align regions having highest similarities b/w 2 sequences
- stretches of sequences with highest density of matches are aligned
- more suitable for partially similar, different length and conserve regions containing sequences.
- Align substring of target sequence with substrings of query sequence
- Suitable for divergent sequences
- Most common ^{local alignment} algorithm.

matrix preparation = negative values → 0

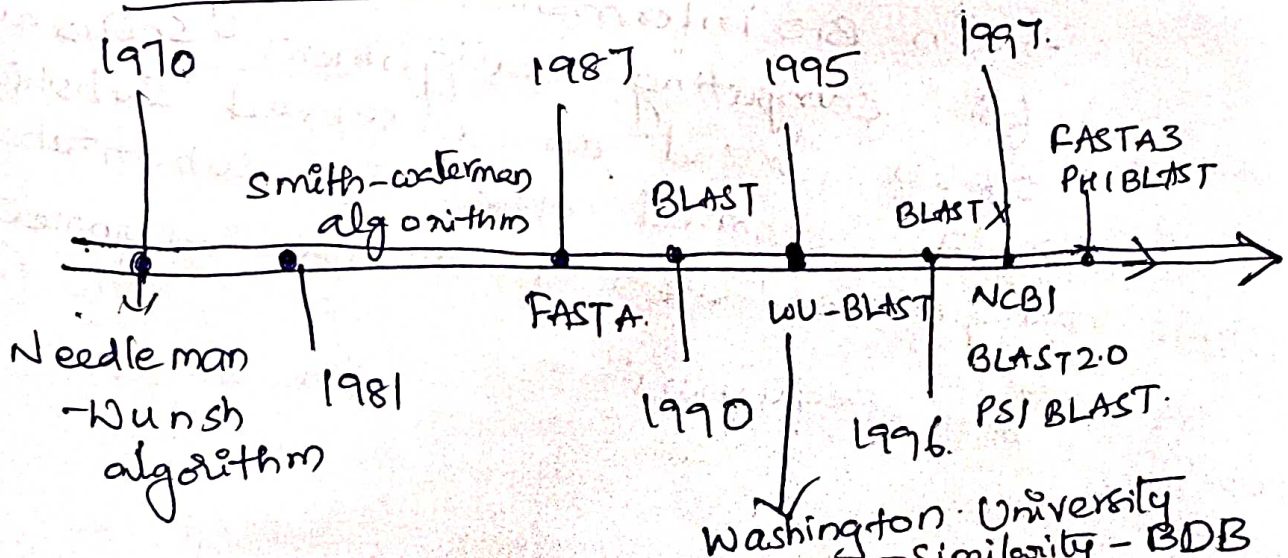
Match → 1

Mismatch → -1

Gap → -2

→ Alignment of short sequence.

History of Alignment algorithms



> BLAST X

NCBI BLAST 2.0 PSI BLAST

→ Position specific, relative protein, evolutionary relationship of protein

PASTA 3 PHI BLAST

→ Search regular expression - local alignment

MEGA BLAST

→ mega sequence alignment, more faster

Scoring - matrix

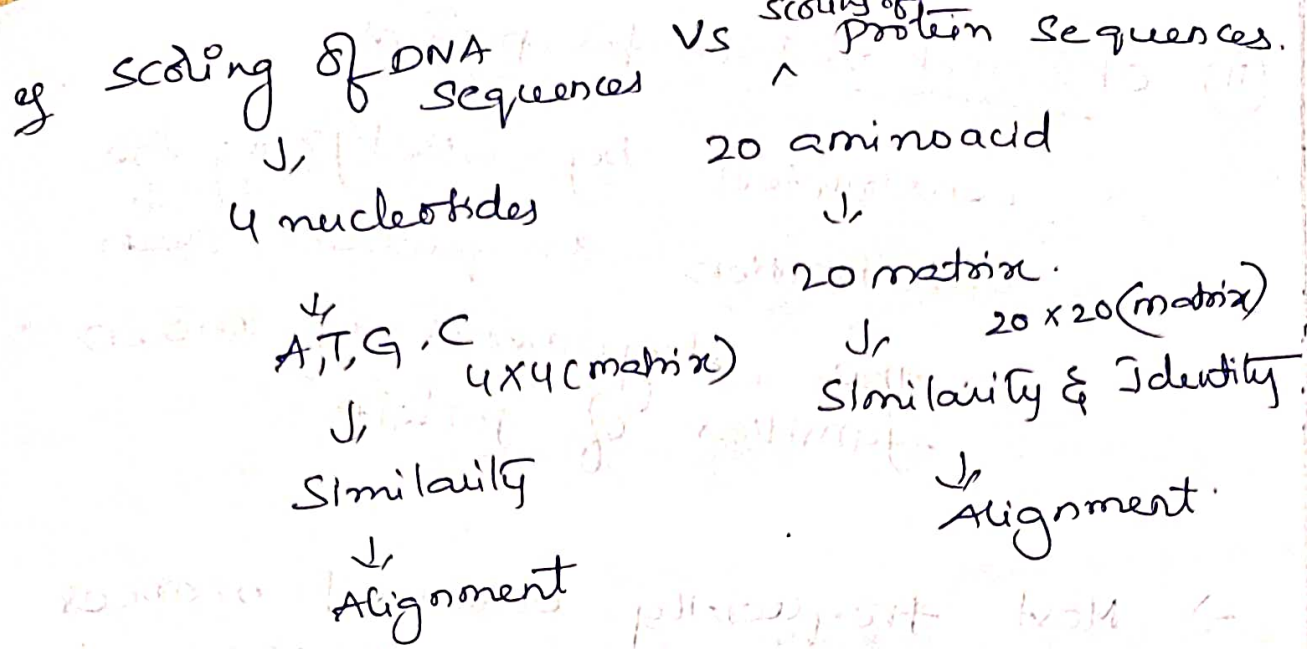
→ Sequence alignment to determine similarity between sequences.

→ One sequence may show similarity with many sequences and to determine most similar sequence among them.

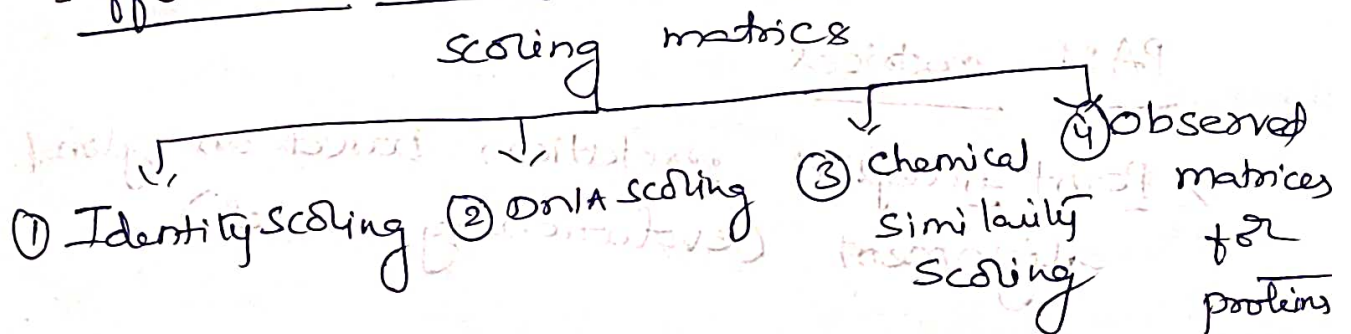
→ we must provide a score to each similar sequence and this score is provided using scoring matrix.

Definition

- In Bioinformatics scoring matrices for computing alignment scores are often based on observed substitution rate, derived from the substitution frequencies seen in multiple alignment of sequences.



Different types of scoring matrices:-



① Identity scoring

— simple scoring where characters are classified as identical (score 1) & Non-identical. — 0 (score).

② DNA scoring

— considers changes as transition & transversion.

scores as
 Identical bp. — 3
 Transition — 2
 transversion — 0.

③ Chemical similarity scoring (for proteins)

→ It gives greater weight to similar chemical properties. (size, shape, charge on amino acid)

→ Structural similarity & opposite.

④ Observed matrices for proteins

→ constructed by analyzing the substitution frequencies seen in the alignments of known families of protein

→ Most frequently observed matrices used are PAM and BLOsum matrices

PAM matrices

→ Point Accepted mutation based on global alignment (evolutionary model).

BLOsum matrices

→ Block substitutions based on local alignments (similarity among conserved sequences).

PAM

→ 1st given by Margaret Dayhoff who compiled alignment of TI groups of very closely related protein sequences.

→ PAM matrices were derived based on divergence between evolutionary sequences of protein structure.

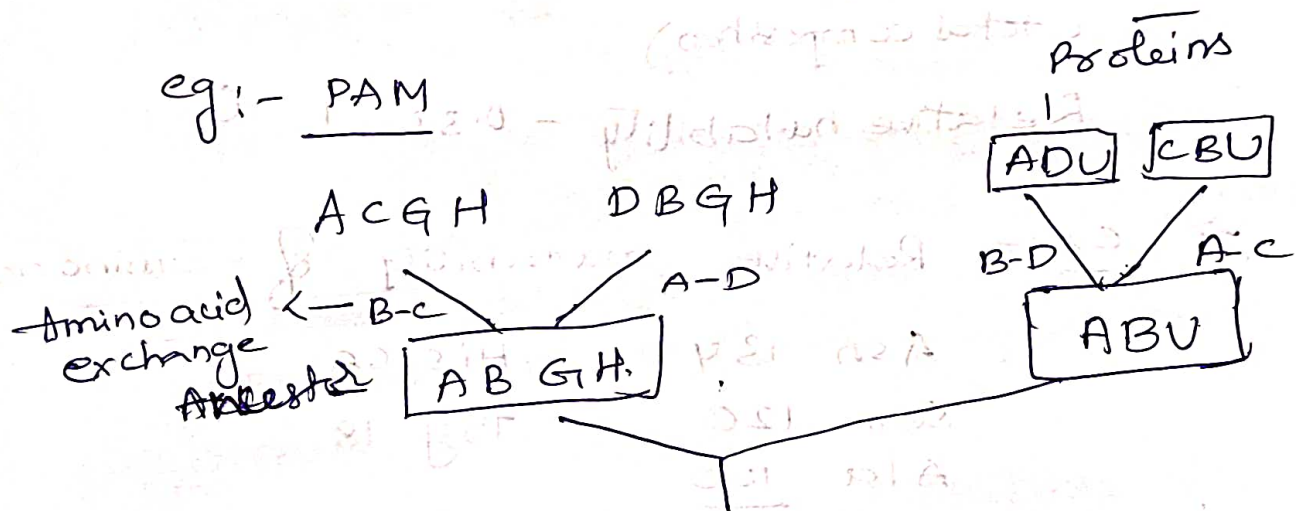
- construction of PAM matrix involves alignment of full length sequence & subsequent construction of phylogenetic trees using likelihood principle.
- PAM accepted by the natural selection.
- A PAM is defined as 1% amino acid change or one mutation per 100 residues.

PAM 1 — 1 substitution per 100 residues

PAM 40 — 40 substitution per 100 residues

- The increasing PAM numbers correlate with increasing PAM units & thus evolutionary distances of protein sequences

eg:- PAM



Phylogenetic tree.

Matrix

	A	B	C	D	E	F	G	H
A			1	1				
B			1	1				
C		1	1					
D		1	1					
E								
F								
G								
H								

Matrix of accepted point mutations derived from the

mutability of Amino acids

→ Relative mutability of amino acid is proportional to the ratio of changes to occurrence.

eg:- A aligned sequence
 ADA - ①
 ADB - ②

Amino acids — A B D
 changes — 1 1 0

Frequency of occurrence
 (total composition) — 3 1 2

Relative mutability — 0.33 1 0

→ ex:- Relative mutability of amino acids

Asn 134 His 66
 Ser 120 Try 18
 Ala 100

Named
 blue

Formula for PAM matrix

① For the Non diagonal values of matrix

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_i A_{ij}}$$

A_{ij} = ~~the~~ element accepted point mutability matrix

λm_j = proportionality constant

$\sum_i M_{ij}$ = mutability of

② $M_{ij} = 1 - \lambda m_j$ - Diagonal values of matrix.

construction of PAM matrix

- for each amino acid type the frequency by which the amino acid is substituted by similar amino acid is calculated
- calculate relative mutability
- calculate mutation probability
- These mutation probabilities are divided by frequency of occurrence of residue
- calculate log of these values
- Completion of matrix.

BLOSUM (BLOCK Substitution Matrices)

- Given by Torja G. Henikoff & Steven Henikoff
- Derived on the basis of direct observation for every possible amino acid substitution in multiple sequence alignment.

- BLOSUM matrix are actual % values of sequence selected for construction of * matrix.

- BLOSUM 62 indicates that sequence selected for constructing the matrix is an average share of 62%.

- BLOSUM share for a particular residue pair is derived from the log ratio of observed residue substitution versus the expected probability of particular residue.

→ BLOSUM γ is the matrix built from blocks with no more than $\gamma\%$ of similarity.

Note: BLOSUM 62 is the default matrix for protein BLAST.

eg: Conserved Blocks Alignment.

✓ ✓

A	A	B	C	D	...	B	B	C	D	A
D	A	B	C	D		B	B	C	B	B
B	A	B	C	D		B	C	C	B	A
A	A	A	C	D						

Construction of BLOSUM γ

- Elimination of sequences that are more than $\gamma\%$ identical.
- The elimination can be done by two methods:

① By removing sequences from the blocks.

② Finding a cluster of similar sequences and replacing it by a new sequence that represents the cluster.

Difference b/w PAM & BLOSUM

PAM

(vs)

BLOSUM

- Based on mutational model of evolution
 - PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.
 - Designed to track the evolutionary origins
 - Higher the number of PAM more divergent species are present
- Based on the multiple alignment of blocks (Local alignment)
 - BLOSUM 62 is a matrix calculated from comparisons of sequences with no less than 62% divergence
 - Designed to find proteins conserved domain
 - Lower the number of BLOSUM more divergent species are present

BLAST

- Basic Local Alignment Search Tool
- It is used to specific scoring matrices (PAM, BLOSSUM) — sequence similarity searches against a variety of sequence databases, to give us high-scoring ungapped segments among related sequences.
- BLAST algorithm is fast, accurate, and web-accessible.
- Faster tool for Sequence similarity
- Provides us with ability to perform analysis by different types of programs

→ Types of BLAST Algorithm

<u>program</u>	<u>Input</u>	<u>Query search</u>	<u>Database</u>
blastn	DNA	1 →	DNA
blastp	protein	1 →	protein
blastx	DNA	6 →	protein
t blastn	protein	6 →	DNA
t blastx	DNA	36 →	DNA

History

- Developed by Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers & David J. Lipman at NCBI

in 1990

- It helps to find regions of local similarity b/w sequences.
- BLAST can be used to functional & evolutionary relationships b/w sequences as well as help identify members of gene families

Types of BLAST

① BLASTN ② BLASTP ③ BLASTX ④ tBLASTn ⑤ tBLASTx

① BLASTN :

- Compares nucleotide query sequences to a nucleotide sequence or a database of nucleotide sequence.

② BLASTP :

- Compares one or more protein query sequences to a subject protein sequence or a database of protein sequences.

③ BLASTX :

- compares a nucleotide query sequence (translated in protein sequence) against a database of protein sequences.

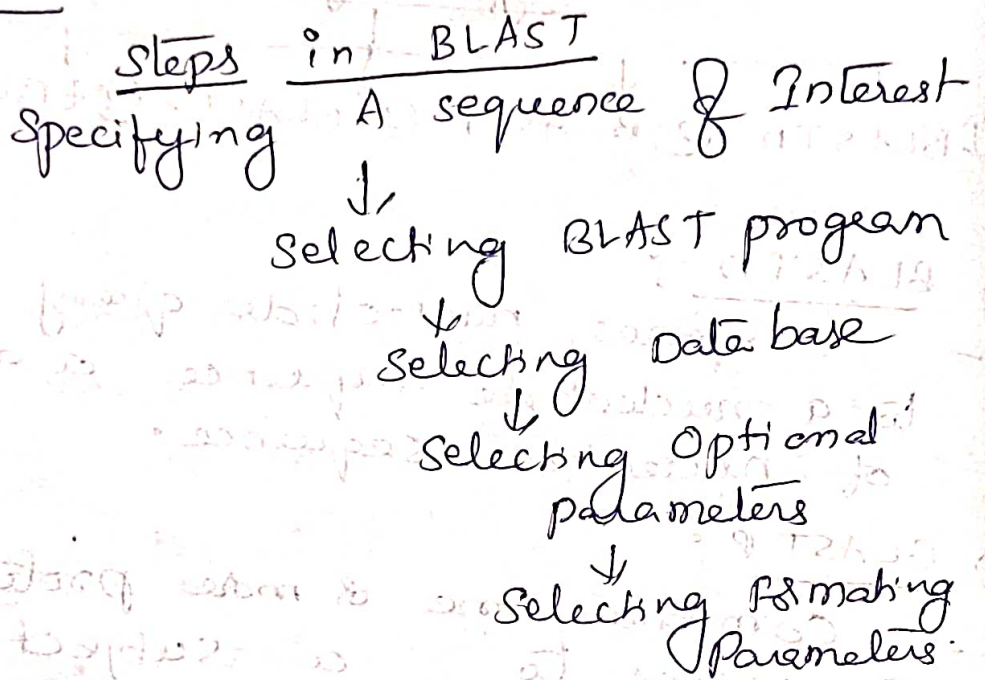
④ tBLASTn :

compares a protein query sequence against the six-frame translation of a database of nucleotide sequences.

⑤ tblast x:

- Compares the translated nucleotide query sequence against the translated nucleotide sequence data base.

Process



Process

query sequence



(Short sequence)

words (specific length)



Three characters words



Compared against
a sequence in a
data base



performed for all words



High scoring
segment pair
(HSPs)

- For each alignment reported an Expect(e) value is reported.

BLAST OUTPUT

→ 3 ways & 3 formats.

- a) Graphical display: Shows where the query, ranked by similarity.
- b) Hit list: number of sequences similar to query, ranked by similarity.
- c) Alignment: every alignment between the query and the reported hits.

Applications of BLAST

- ① Homology Searching — similarity
- ② Species Identification — comparison of species
- ③ Establishing phylogenetic relationships.
→ query sequence — match with similar sequence
- ④ DNA mapping, comparison of sequences.

Unit - 4

Molecular Evolution

- Molecular Evolution is the process of change in the sequence composition of cellular molecules such as DNA, RNA and protein across generations.
- Molecular evolution uses principles of evolutionary biology and population genetics to explain patterns in these changes.

History of molecular evolution

- It started early 20th century with "Comparative biochemistry"
- molecular evolution came - 1960s & 1970s.
- Advent of protein sequencing allowed molecular biologists to create phylogenies based on sequence comparison and to use the differences b/w homologous sequences as a molecular clock to estimate the time since the

last common ancestor

→ In the late 1960s, the Neutral Theory of molecular evolution provided a theoretical basis for the molecular clock.

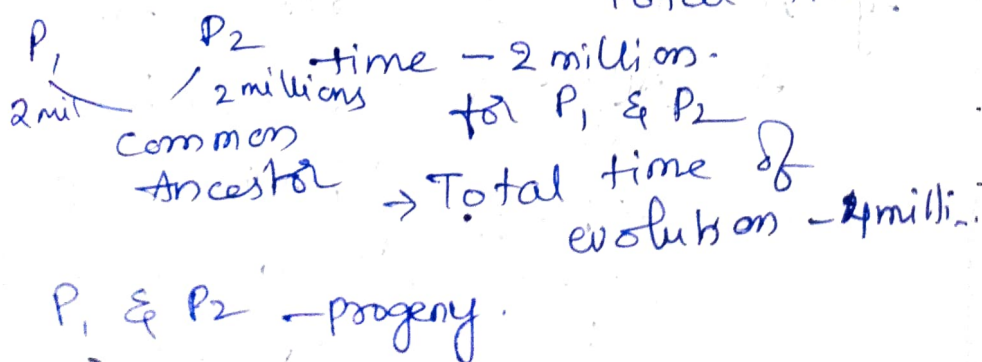
→ After 1970s, nucleic acid sequencing allowed molecular evolution to reach beyond proteins to highly conserved ribosomal RNA sequences, the foundation of a reconceptualization of early history of life.

Rate of molecular evolution

- The rate of evolution is estimated from the amino acid sequences of a protein & nucleotide sequences in a region of DNA in two or more species

$$\text{rate of molecular evolution} = \frac{\text{No of amino acids} + \text{No of nucleotide substitution}}{\text{Total time}}$$

eg:- Humans



phylogenetic trees

Scaled
Trees

Branched lengths
are equal to
the magnitude
of change
in the nodes.

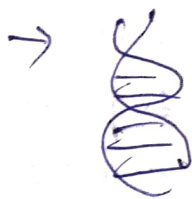
unscaled Trees

— only representing
the relationship
B/w sequences



cellular memory
unit

→ Translated
product
(protein)



DNA

→ modified

mutation
Substitution
Insertion
Deletion

molecular
evolution

→ Mutation: Sudden change in DNA or RNA

→ Substitution: when DNA sequences diverge, they begin to

eg: A G T C - seq ① collect mutations.

A G A (G) - seq ②

↓
Substitution
c-replaced
by G.

The number of
Substitutions (P) found
in an alignment is
widely used in molecular
evolution analysis.

→ Insertion :- Addition of nucleotide or addition of aminoacid.

→ Deletion :- Removing of nucleotide chain or removing of aminoacid.

→ Neutral mutation :- mutation has no fitness effects.
Neutral theory invisible selection.

→ Motoo Kimura, Thomas Jukes
(1968, 69)

→ concept of molecular clock

→ Most of the aminoacid changes & polymorphisms are neutral and created by genetic drift & mutation.

evidences of neutral evolution

① Pseudogenes (dead genes that have no function and no fitness effect) evolve very fast.

② Genes that have important functions should evolve slower.

③ Synonymous codon positions (3-fold, 4 fold degenerate sites) evolve faster than non-synonymous sites should evolve with constant rate.

phylogenetic tree :-

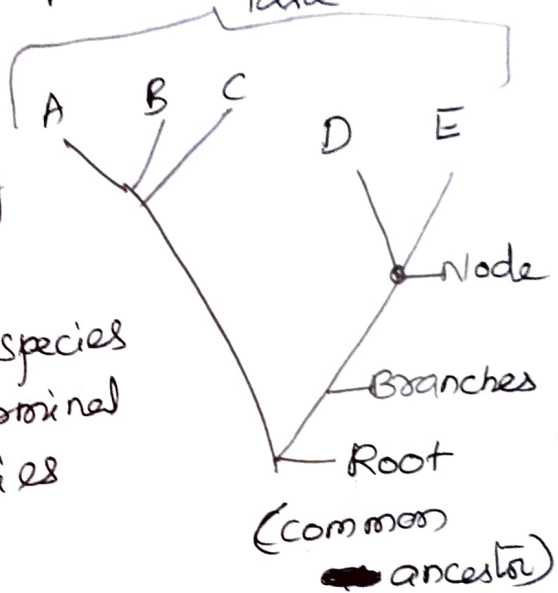
phylogenetic tree is a diagrammatic representation of evolutionary relationships among living organisms

→ It represents how different species evolved from a series of common ancestors

→ The phylogenetic tree is also called the "Tree of Life" or "Dendrogram"

Node = ancestor of present day species

Taxa = present day species
~~off~~ Terminal species



Types of phylogenetic trees

① Rooted tree

② Unrooted tree

① Bifurcating tree

② multifurcating trees

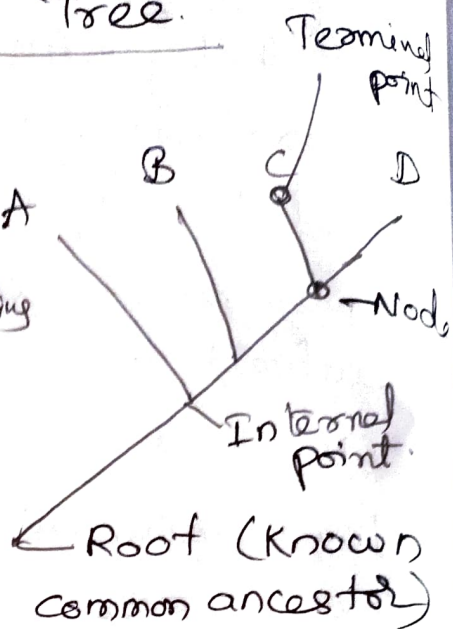
Bifurcating tree

multifurcating trees

① Rooted phylogenetic Tree.

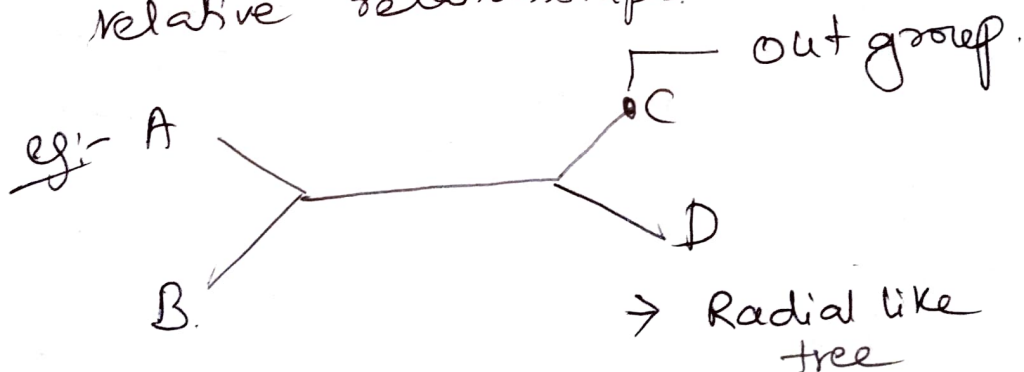
→ It is directed tree with a unique node A

- the root - corresponding to the most recent common ancestor of all the entities at the leaves of tree.



② Unrooted phylogenetic tree.

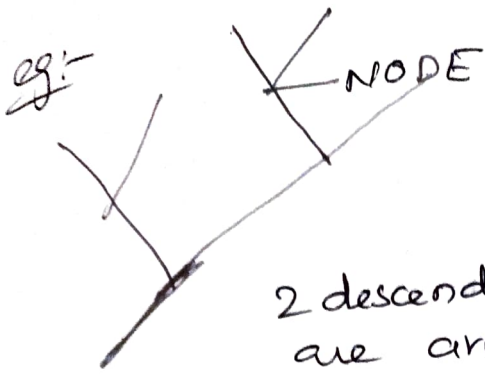
- An unrooted phylogenetic tree does not give the information of a common ancestor, but only positions the taxa to show their relative relationships.



a) Bifurcating phylogenetic Tree.

- A bifurcating tree has exactly two descendants arising from each interior node. Both rooted and

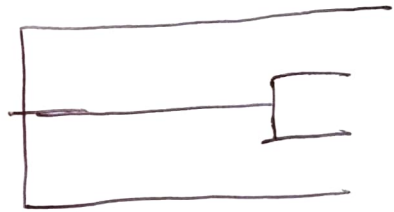
unrooted trees can be bifurcating.



2 descending
are arising - bifurcating

b) multifurcating phylogenetic tree

- multifurcating tree has multiple descendents arising from each of the interior node. Both rooted and unrooted trees can be multifurcating



Importance of phylogenetic tree.

- To study evolutionary histories
→ How the species were spread geographically.

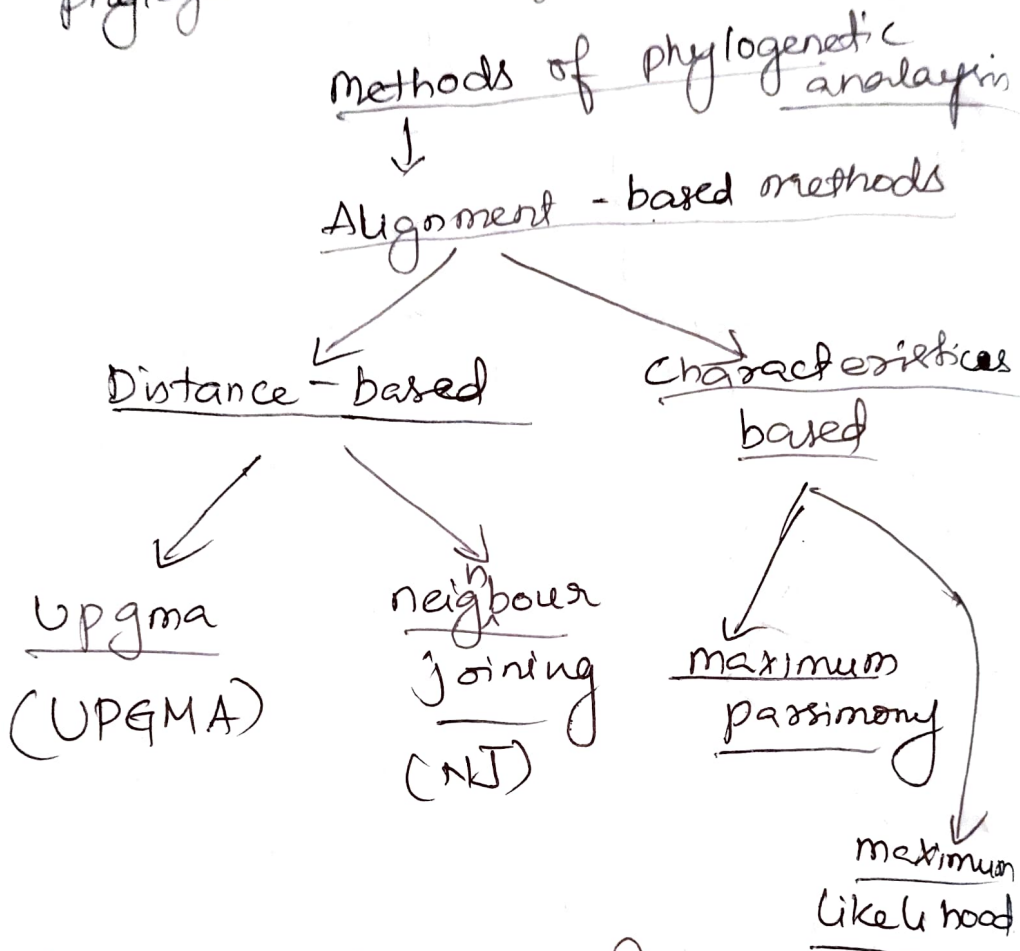
→ Search for a new species

→ To identify the most recent common ancestor.

→ microbes can be traced along with their evolutionary histories.

Methods of phylogenetic Analysis

→ Alignment - Based methods of phylogenetic analysis.



→ Distance - based methods of phylogenetic

→ Use the dissimilarity (the distance) b/w two sequences to construct trees.

→ They are much less computationally intensive than the character based methods are mostly accurate as they take mutations into count

→ for tree generation, generally hierarchical clustering

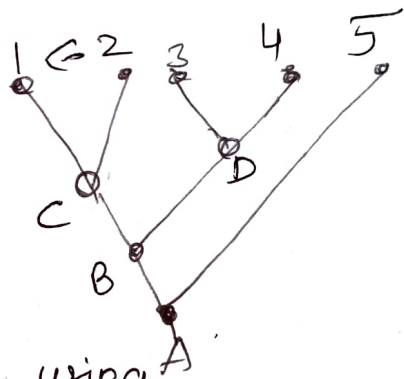
is used in which dendrograms (clusters) are created.

a) UPGMA :

(Unweighted Pair Group Method with Arithmetic Mean)

- A simple clustering method that assumes a constant rate of evolution (molecular clock hypothesis).
- It needs a distance matrix of the analysed taxa that can be calculated from a multiple alignment.

eg:-



Steps involved to construct a phylogenetic tree using UPGMA :

- Align & Name them
- Compare sequence them using pair-wise sequence alignment
- Count the mismatches and record them in the table
- Complete the table by comparing all the sequences.
- look for the lowest value to find the first group.

→ By finding the arithmetic mean step by step, and construct the tree.

Example:-

A - ATCGATCG
 B - GTA GACGA
 C - ACCGTACG
 D - TCA GTCAC
 E - CCTACAG

→ Pair-wise sequence alignment by mis match

① A - ATCGATCG mismatch - 5
 B - GTA GACGA sequences

② A - ATCGATCG
 C - ACCGTACG
 → 3

	A	B	C	D	E
A		5	3	6	3
B			7	5	6
C				4	5
D					4
E					

Find lowest value - ③

A
 └─ C.

1st group.

	A/C	B	D	E
A/C		6	<u>5</u>	5
B			5	6
D				<u>4</u>
E				

lowest value

$$① \frac{(A-B) + (C-B)}{2} = \frac{5+7}{2} = \frac{12}{2} = 6$$

$$② \frac{(A-D) + (C-D)}{2} = \frac{6+4}{2} = \frac{10}{2} = 5$$

Average 1st group

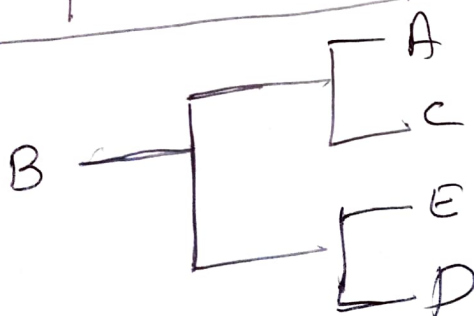
second group

	A/C	D/E	B
A/C		5	6
D/E			<u>5.5</u>
B			

lowest value

$$\frac{A}{C} = \frac{[A/C-E] + [A/C-D]}{2}$$

$$= \frac{5+5}{2} = \frac{10}{2} = 5$$



Phylogenetic Tree

$$D/E \& B = \frac{D-B + E-B}{2} = \frac{5+6}{2} = 5.5$$

b) Neighbour joining (NJ)

- bottom-up clustering method that also needs a distance matrix
- NJ is a heuristic approach that does not guarantee to find the perfect result, but under normal conditions has a very high probability to do so. It has a very good computational efficiency making it well suited for large datasets.
 - constant rate can't be assumed
 - mutation rate - Base pair
 - Branch length is proportional to rate of evolution

2) characteristics based methods

- character based methods compare:
- All sequences simultaneously considering one character/site at a time
- These are maximum parsimony and maximum likelihood
- These methods use probability

and consider Variation in a set of sequences.

a) Maximum Parsimony (MP)

→ This method tries to create a phylogeny that requires the least evolutionary change.

→ It may suffer from long branch attraction, a problem that leads to incorrect trees in rapidly evolving lineages.

b) Maximum - Likelihood (ML)

→ ML uses a statistical approach to infer a phylogenetic tree.

→ ML is well suited for the analysis of distantly related sequences, but is computationally expensive and thus not that well suited for larger input data.

Unit - 5

Genomics

Genomics

→ Gene - gene

→ omics - study in biology ending in the suffix - omic.

→ Genome - Hans Winkler (1920)
Coined the term

→ Sanger (1970s) - Genome - fundamental part of many omics
↑
Sequencing of genome.

→ Word genomics is used in 1980's
widely used in 1990's.

→ Genome - genetic material of an organism.

↑
Coding genes

Non coding genes.

{ contains DNA genes
and Mitochondrial
& Chloroplast DNA.

→ Genomics - study of genome.

eg:- 46 chromosomes of humans
22 pairs - (Autosomes) - Somatic chromosomes
1 pair - Sex chromosomes (Allosomes)
46 chromosomes

↓
Study of genes in 46

Chromosomes - human genome.

- cells in a human body
↓
100 trillion
- chromosomes in a human cell
↓
46
- Genes in a human genome
↓
20,500
- DNA is - 2 mls long.
- Genomics uses
- Recombinant DNA,
 - DNA sequencing method
 - Bioinformatics -
to sequence, assemble,
and analyze the structure
function of genome.
- ① human genome in humans
↓
99.9% same
- ② 1% of the genome contains genes
↓
2%
- ③ 1% of human proteins that share similarities with other organisms
↓
50%

Sequencing technologies:

- ① Sanger sequencing (1st generation sequencing)
- ② Next-generation sequencing (NGS)

① Sanger Sequencing:

- ① Fredrick Sanger with his colleagues developed (1977). - won Nobel prize in 1980.

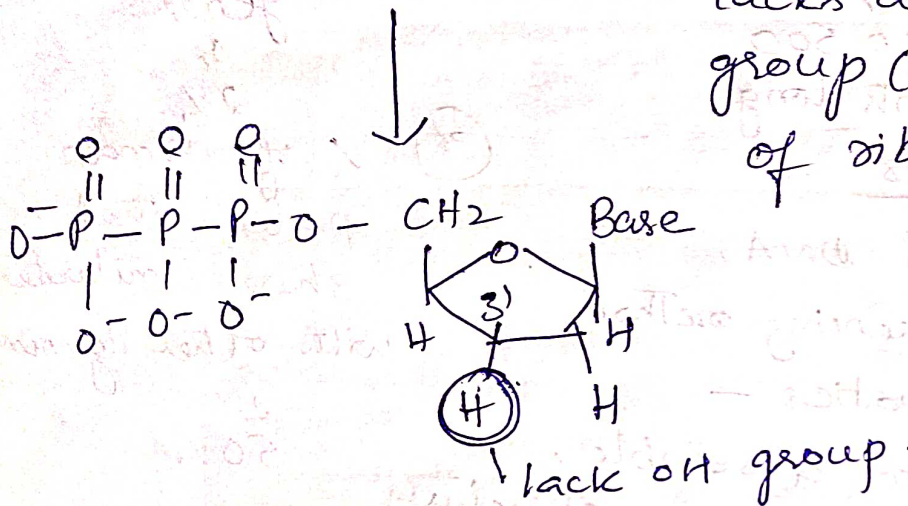
- Also called chain termination method,
(Dideoxy sequencing (DDNA Sequencing),
Enzymatic method)

- used in Human Genome Project
(2003)

→ These method is based on the chain termination by the use of ddNTPs ~~Di~~ dideoxynucleosides (ddNTPs)

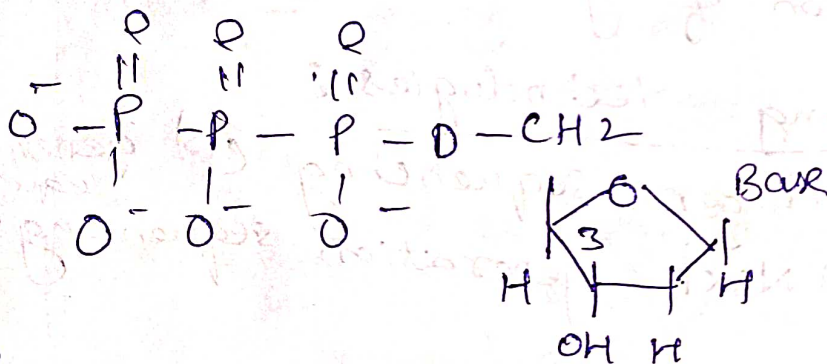
ddNTPs

→ A ddNTPs — Artificial molecule lacks a hydroxyl group (OH) at 3' carbon of ribose ~~sugar~~ sugar



↓
can't bind to nucleotide.

dNTPs



Requirements for Sanger sequencing

① DNA

② Primer

③ DNA polymerase

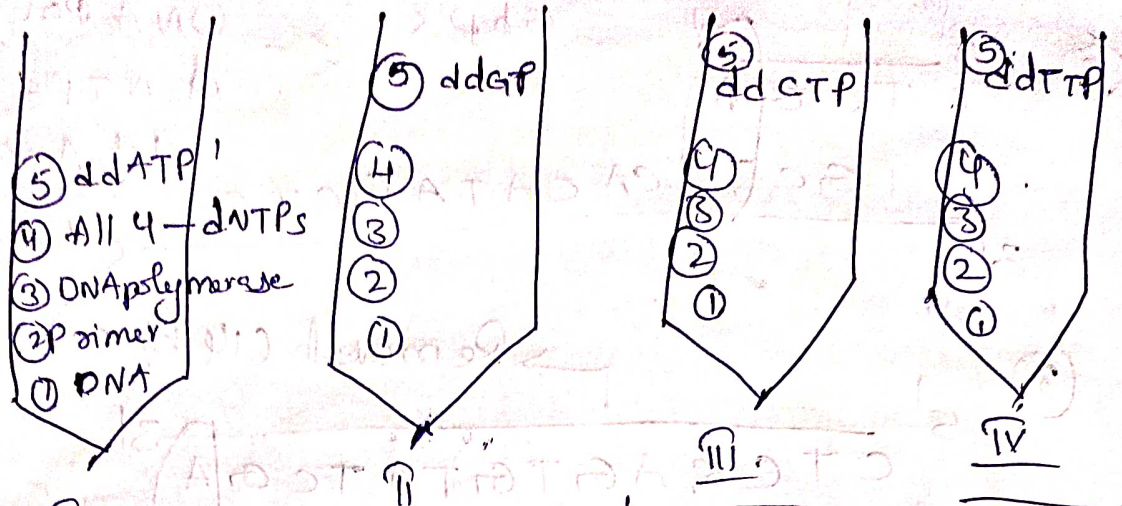
④ dNTP's [Deoxyribonucleoside

⑤ ddNTPs

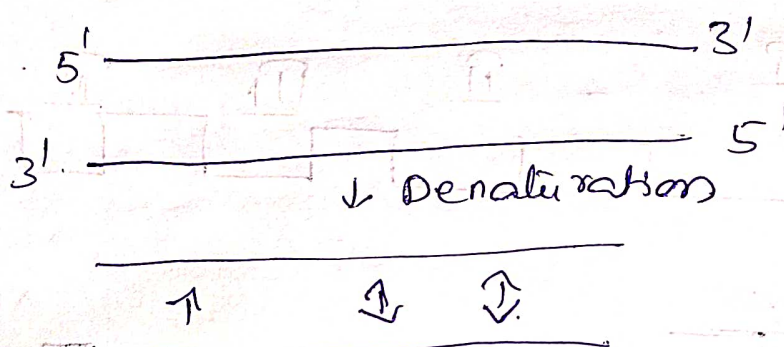
[dideoxyribonucleoside
Triphosphate]

dATP
dGTP
dCTP
dTTP

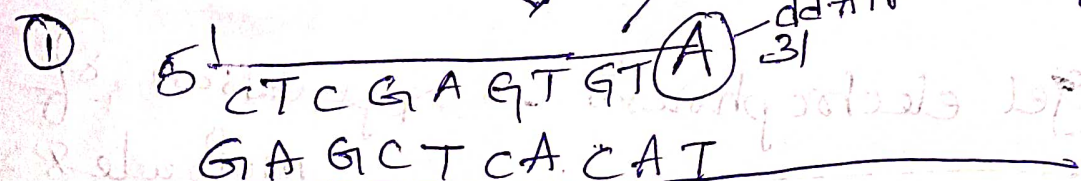
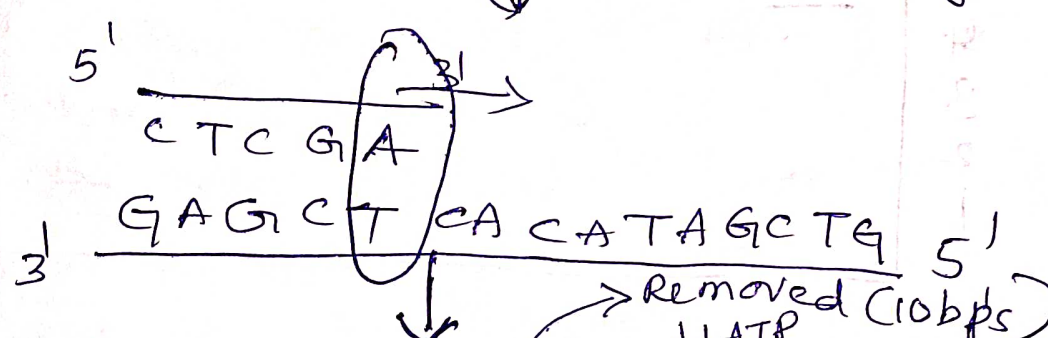
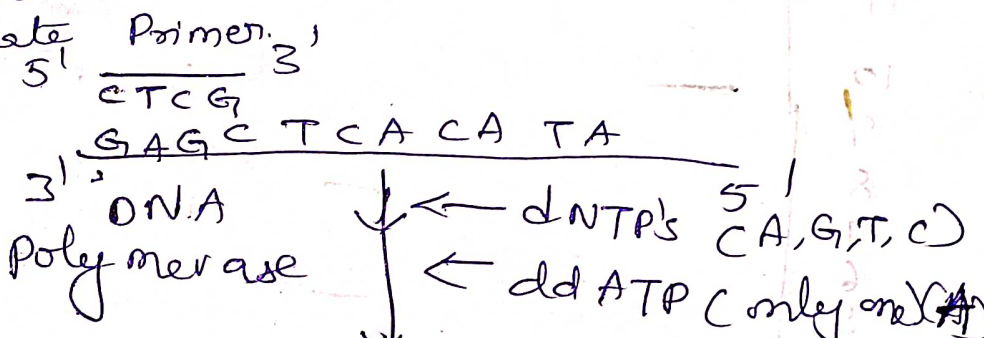
ddATP
ddGTP
ddCTP
ddTTP

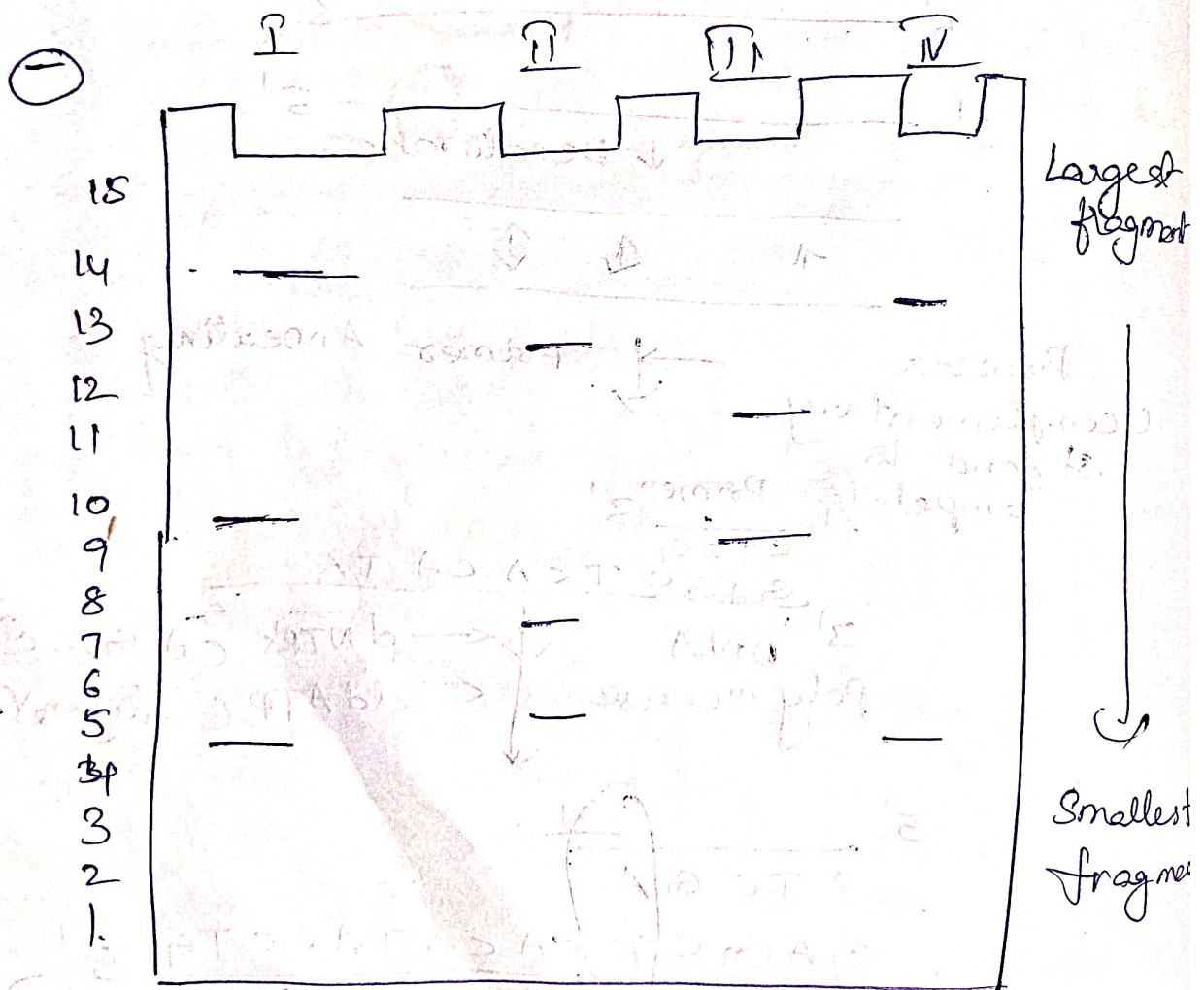
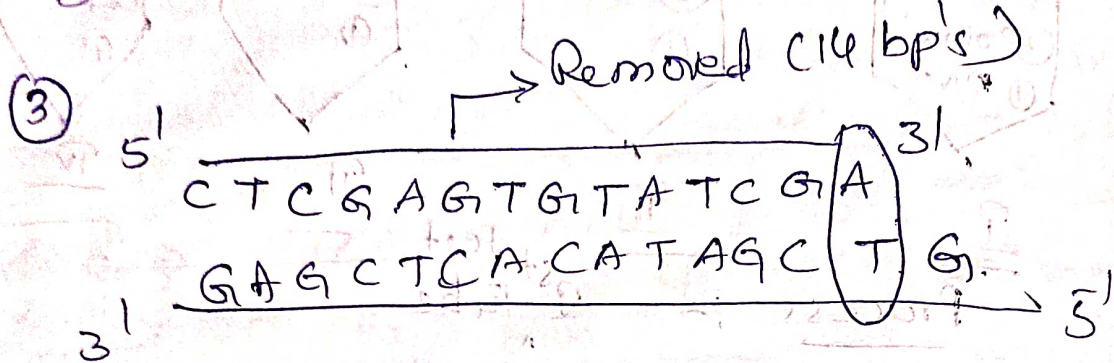
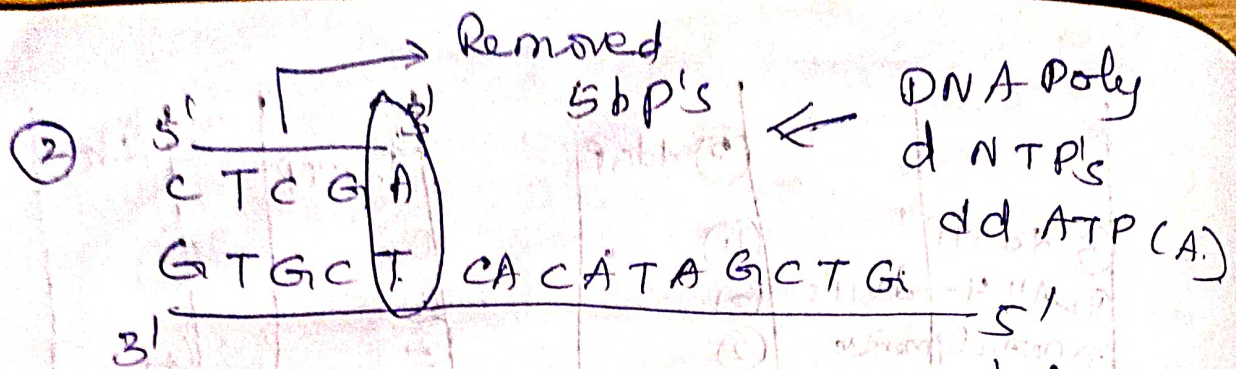


Process :- 1st test tube



Primer complementary strand to template → primer Annealing

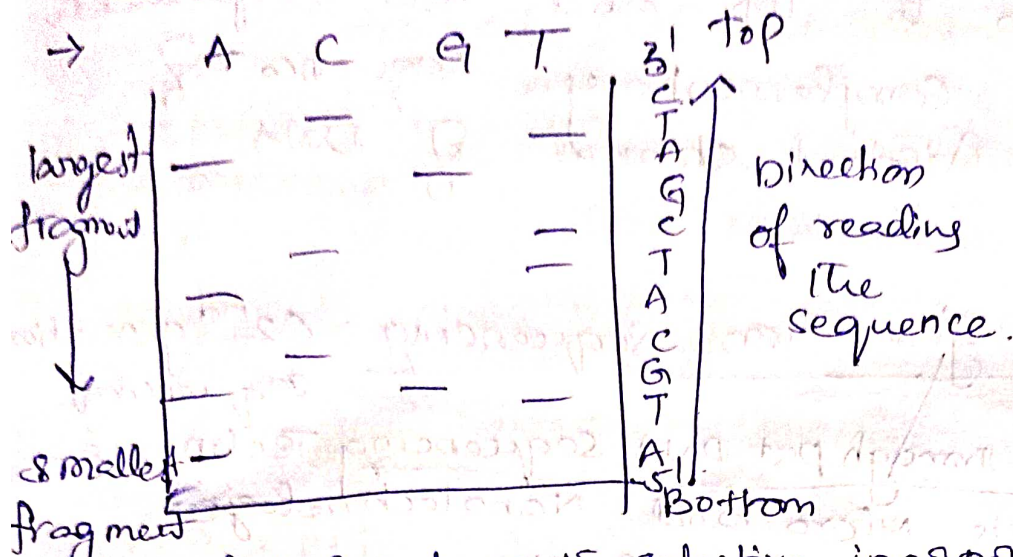




→ Gel electrophoresis — Separation of molecules

→ Agarose gel electrophoresis → Polyacrylamide gel electrophoresis

Autoradiography



Principle: Based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during *in vitro* DNA replication.

Steps:

- ① A primer is annealed to a single-stranded section of DNA.
- ② DNA-primer mixture is put into 4 separate tubes with DNA polymerase and a solution of dNTPs at a concentration of 100 times lower than the dNTP concentration.
- ③ DNA pol uses dNTPs to extend the DNA.
- ④ ddNTPs are put together randomly, resulting in different lengths of fragments.
- ⑤ Fragments that are from each of the reactions are denatured and separated by size using gel electrophoresis.
- ⑥ The gel is used to visually detect the DNA fragments.

fragments are to be read from bottom to top and this represents the complementary sequence of the original strand of DNA.

② Next generation sequencing (2nd generation sequencing)

→ High throughput DNA sequencing technique

→ Employs Micro and Nanotechnologies:

a) Reduce Sample size.

b) Low Reagent cost

c) Less Time.

→ Massive parallel sequencing

→ Sequence thousands of sequences at once.

→ Produce enormous amount of data.

→ NGS Techniques.

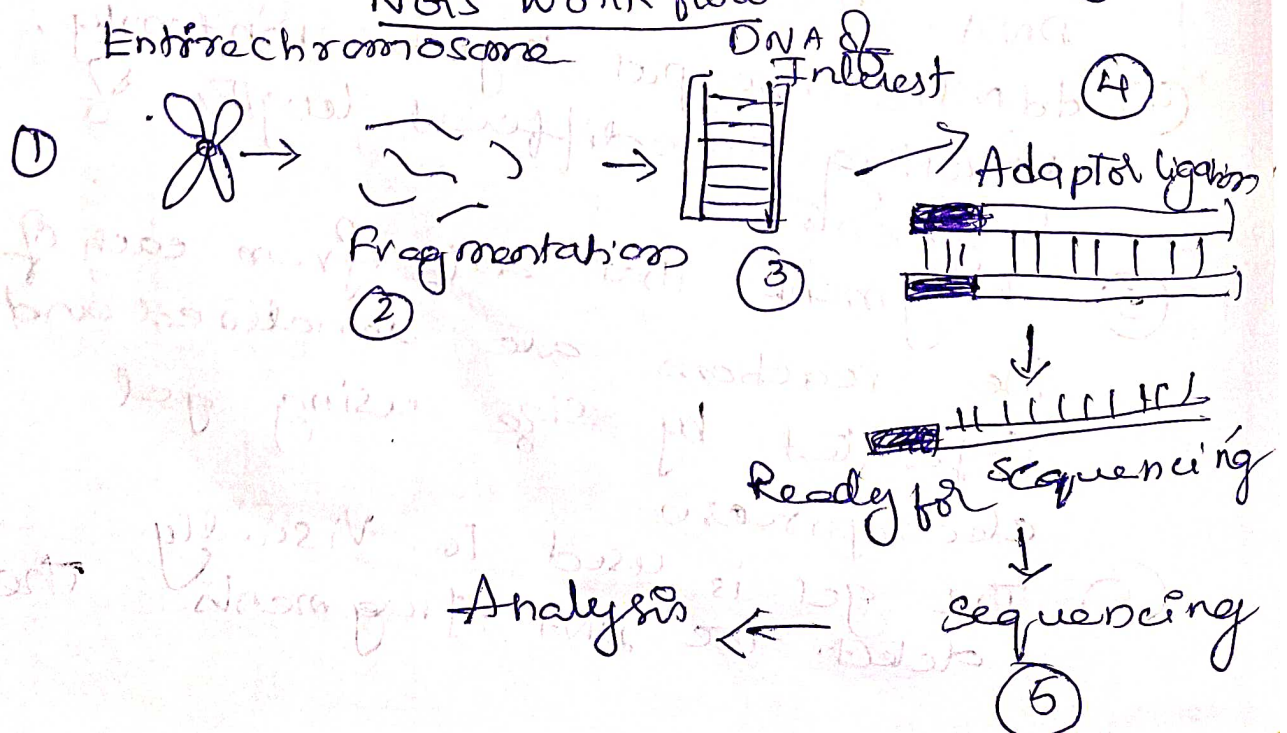
① Illumina sequencing

② Pyro-sequencing

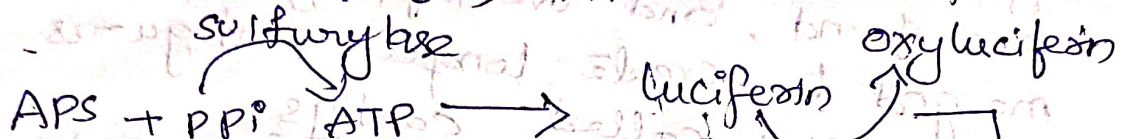
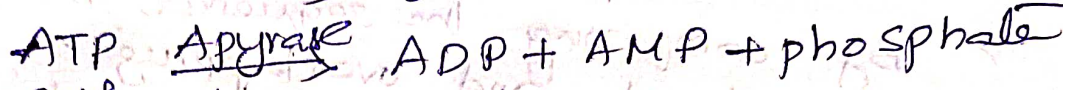
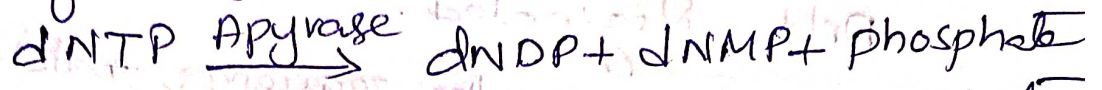
③ Roche 454 sequencing

④ Ion-torrent sequencing.

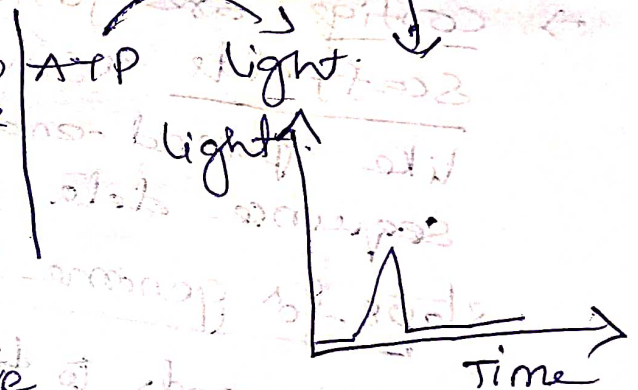
NGS work flow



① Pyrosequencing: Non-electrophoretic, bioluminescence method that measures the release of inorganic pyrophosphate by proportionally converting it into visible light using a series of enzymatic reactions.



→ Nucleotide Incorporation generates light seen as a peak in the Program trace.



Key principles

- ① DNA sequencing: Libraries are generated by clonal amplification by PCR in vitro.
- ② DNA is sequenced by synthesis: Addition of Nucleotides
- ③ sequencing: Amplified DNA templates are sequenced simultaneously in a massively parallel fashion without the requirement for a physical separation step.

Steps involved in NGS

- ① Nucleic acid Extraction & Isolation
- ② Library Preparation
- ③ Clonal Amplification & sequencing
- ④ Data Analysis using Bioinformatics

Genome Assembly

→ The process of putting a large number of short DNA sequences back together to recreate the original chromosomes from which the DNA originated.

→ By comparing the sequences, matching regions are found, and overlapping reads are merged to create longer contiguous sequences called contigs.

→ Contigs are further assembled into larger scaffolds using additional information like paired-end reads or long-range sequence data.

Steps for genome assembly

- 1) Align reads to find overlapping regions
- 2) Determine a consensus sequence (or contig)
- 3) Scaffold contigs based on read pairs and/or overlapping regions
- 4) Generate pseudogenes based on genetic maps

DNA
↓
sequences



contig



scaffold.



Mapped
scaffolds.

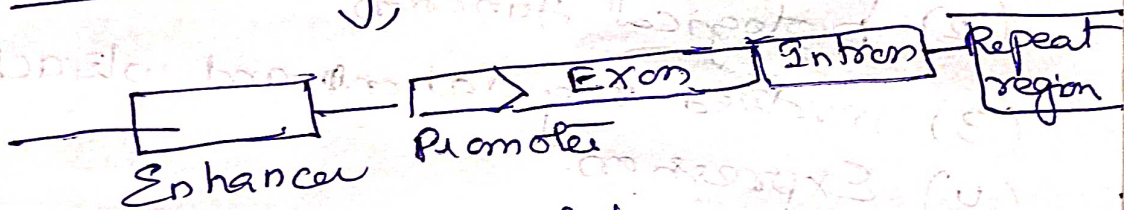
Genome annotation :- Dr. Owen White (1995)

→ It is the process of identifying the location of genes and discovering their function, as well as annotating other genetic features from a raw genome sequence.

→ Genome sequence

CATGACGTGCTACCA

↓
Genome Annotation



Genome Annotation

↓
Structural Annotation

→ Structure prediction

→ Identifying elements

→ Introns

→ exons, stop, start codons.

↓
Functional Annotation

→ function prediction

→ contains Biological information
↓
which protein is produced

Structural Annotation

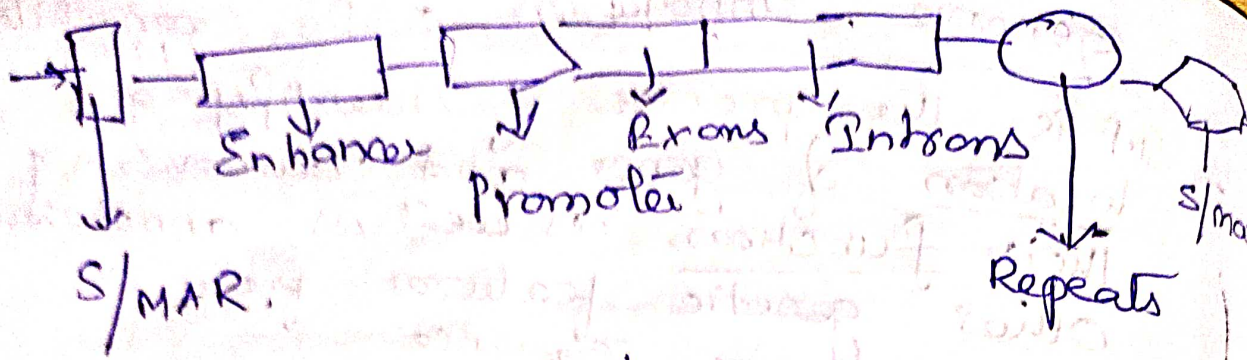
- identification of genomic elements

→ Open reading frame and their ~~localisation~~ localisation

→ gene structure

→ coding regions

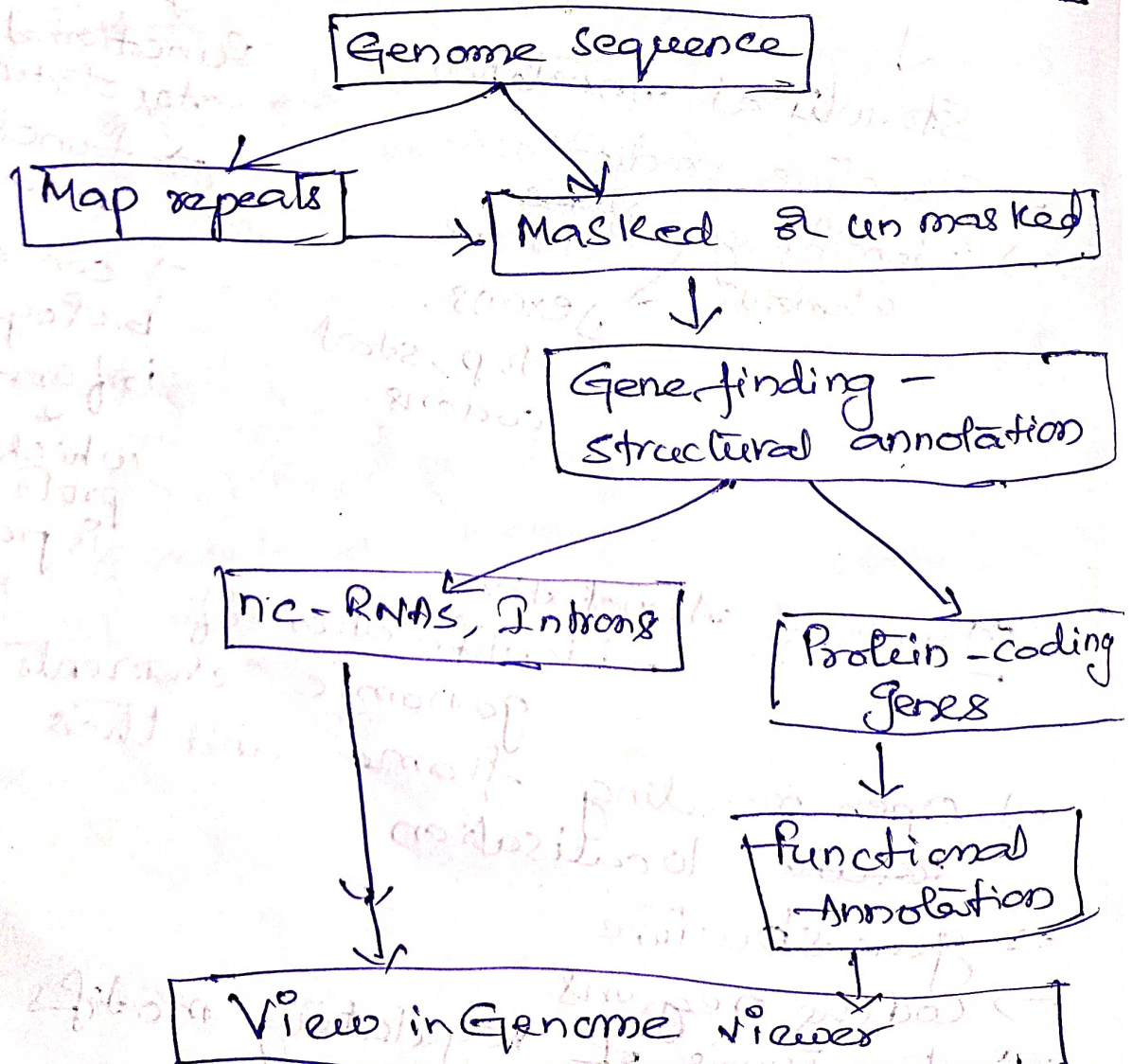
→ locations of regulatory motifs.



Functional annotation

- Attaching biological information to genomic elements.
- (1) biochemical function
 - (2) biological function
 - (3) involved regulations and interactions
 - (4) Expression.

Genome Annotation workflow



Genome of ~~Arabid~~ Arabidopsis (2M)

- Arabidopsis — 2000
Genome project (1990-2000)
 - ↳ Cruciferae
 - ↳ Rapid growth
 - ↳ low DNA content
 - ↳ small genome size
- contains about 125 megabases of sequence
- Encodes approximately 25,500 genes
- Contains a similar number of gene functional classifications as other sequenced eukaryotic genomes (Drosophila melanogaster & Caenorhabditis elegans)
- has 85% unique genes
- has 39.5% genes that exist as members of large gene families
- shows evidence of ancient polyploidy as estimated 58-60% of the Arabidopsis genome exists as large segmental duplications.
- Analysis of sequence of the Arabidopsis genome → Similar to multicellular organisms.
- It also differs from animals
- Plants are autotrophic → Require light, H_2O , minerals
 - ↓
 - photosynthesis.
- But animal genome differ from plant genome due to ~~plant~~ Heterotrophic mode of nutrition.
- Arabidopsis centromeric regions — heterochromatic
 - ↓
 - 4.7 expressed genes are prot.

- Arabidopsis has unique genes. — approximately 150 unique protein families.
- Arabidopsis genome contains genes encoding RNA polymerase subunits not seen in other eukaryotic gene. 16 unique families of transcription factors.
- Arabidopsis — contains several classes of proteins — used in animal systems.

~~Arabidopsis~~

Difference b/w Sanger sequencing & NGS

Sanger Sequencing

- ① Sequence Volume. → ① low sequence volume
- ② Sensitivity → ② low
- ③ Speed & Cost. → ③ Slow turnaround time with high cost per base.
- ④ Time taken for completion → ④ long time
- ⑤ Sequencing → clones, PCR samples

NGS Sequencing

- ① millions of reactions in parallel, resulting in high-throughput
- ② High.
- ③ Quick ~~turnaround~~ time with low cost per base.
- ④ Short time
- ⑤ DNA libraries

- ⑥ preparation steps → ⑥ few, sequencing reaction clean up Many, Complex procedures
- ⑦ Data collection → ⑦ samples in plates: 96, 384 samples on slides 1-167

Applications of NGS

- Bioinformatic analysis
- NGS → hundreds of megabases to gigabases of nucleotide sequence at a time
- genome sequencing — used in biomedical sciences
- low cost of sequencing

Genome of Mycobacteria

- Mycobacteria — vet gram bacteria
- Mycobacteria — 1882 — Robert Koch
- Robert Koch — discovered along with M. Tuberculosis, cholera vibrio, Bacillus anthracis.
- won nobel prize — 1905

characters

- gram +ve bacteria, 4.4 million base pairs.
- slow growth
- Dormancy
- Complex cell envelope
- Intracellular pathogenesis
- Genetic homogeneity

species

M. africanum — Humans
M. bovis — cattle.
M. microti — voles
M. canettii — Humans
M. tuberculosis — Humans.

Aim :-

- Transmission, clinical ~~phenotypes~~ phenotypes
- develop advanced vaccine against the TB
- develop drug

steps :- Involved in Genome sequencing
Genomic DNA



Cosmid / BAC library



Shotgun sequencing



Genome sequence



Relational Database

→ cosmid — (1.4 - 2.0 kb)

BACs — were cloned into M13 Vectors.

→ less than 0.1% — clone coverage

< 0.2%, from one strand

with sequence chemistry

(Informatics

→ data base — EMBL, TREMBL, SWISS PROT
BLASTN, BLASTX, FASTA

→ M. tuberculosis — 4,411,529 bp

— G+C content — 65.6%
— Genome Size 4.4 Mb (4,411,529 bp)

Unit-6

→ Proteome → coined by Marc Wilkins 1994
→ coined in 1995
→ entire set of proteins

→ Cell (tissue, organism) → proteins produced in certain time & defined conditions.
genome.

→ study of proteome - Proteomics

↓
(all the proteins produced by a cell different conditions)

→
Genome (20000, 25000 genes) → Alternative splicing → Transcriptome (100,000 mRNA transcripts - 1 Lakh) → Proteome (more than 10,000 proteins)

Complexity

→ proteomics helps to understand the character, functions of protein in biological organism.

→ proteomics involves

- Protein purification
- Identification
- Structural analysis
- Interaction
- Function.

→ Technologies

Such as

Immunosay
Mass Spectrometry
are used

Applications

→ Identify proteins involved in pathological processes

→ drug industry include targeted identification and investigations of mechanisms of drug action and its toxicity.

→ Evolutionary studies.

Types of proteomics

- Structural - study of 3D structure
- functional - study of functions.
- Expression Proteomics

↓
study of qualitative & quantitative expression of proteins in different environments such as diseased & normal individual.

Bioinformatics views

- Data bases of protein sequence
- Data bases of predicted protein structures
- Data bases of protein expression analysis

Structural proteomics



Sample → Proteins → peptides.

- studies protein-protein interactions
- Identifies all proteins
- structural analysis of protein.

Functional proteomics

- Identification
- Quantification
- Modification
- Structure
- Function

MASS SPECTROSCOPY

Proteomics - tools

<u>Method</u>	<u>Description</u>	<u>Applications</u>
① Mass Spectrophotometry	Digest protein and fragment peptide to identify proteins	Protein identification, sequence, post-translational modification
② protein chips	Synthesize proteins, peptides, antigens, antibody into a very format and spot onto slides.	Protein interaction with lipid, protein, molecules, drug discovery, post translational modifications
③ Bioinformatics	In silico proteomics	mining database, predicting protein interactions.

Significance of proteomics

- Protein profiling - structure & function
- protein arrays - interactions studies.
- Proteomics to a phosphorylation
- Proteome mining.

Techniques:-

- ① Mass spectrometry (MS2)
- ② NMR spectrometry
- ③ Isoelectric focusing etc.

Proteome Analysis

MS2

Extraction → Pretreatment → Purification → Identification

Ex: Peptidomic
Inactivation
- Heat
- Inhibitors
- Cold acid.

↓
RP-HPLC Quantification

- ① ELISA
- ② RIA

Proteomics → Tryptic digestion → RP-HPLC → MS2
Capillary reversed.

↓
Labeled tryptic peptides.

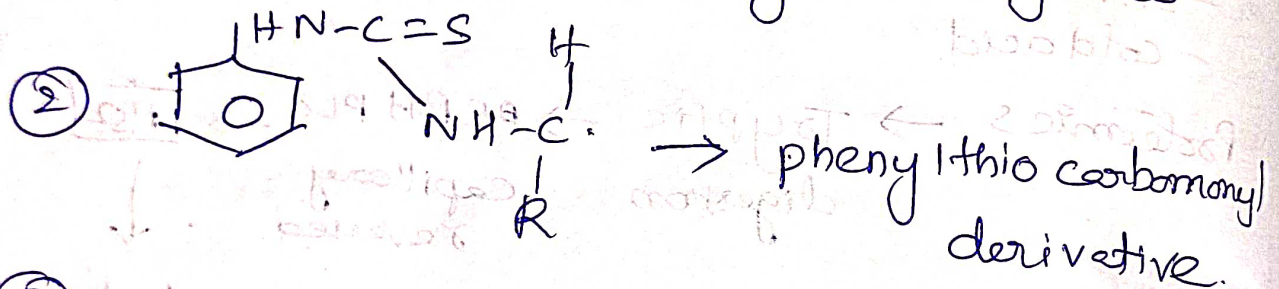
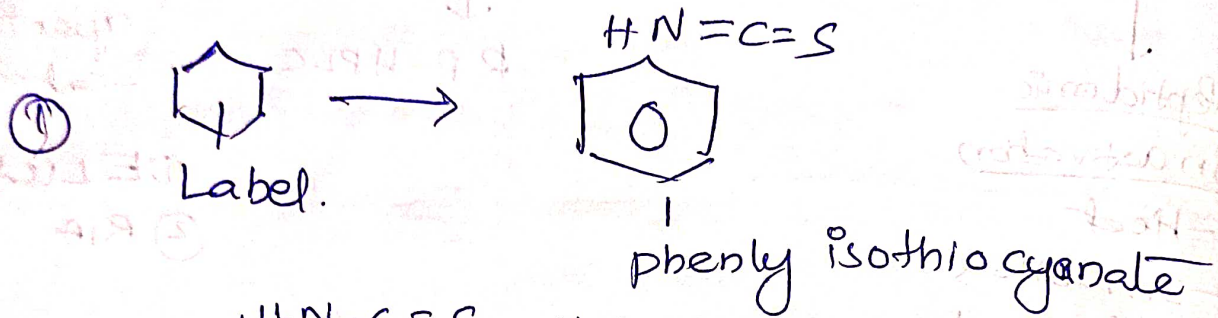
Sequencing techniques of Protein

- ① Edman degradation:- Performed by Edman in 1950
- sequencing of Amino acids in protein

Terminal peptides
N-
C-terminal
Label.
↓
Identify
by chromatography/
electrophoresis.

→ Labelling of N-terminal of protein by phenyl isothiocyanate (Edman's ~~reagent~~ reagent).

→ The reaction b/w Edman's reagent to N-terminal of protein is done in mild Alkaline condition.



③ Phenylthiocarbamoyl derivative $\xrightarrow[\text{Condition.}]{\text{Acidic}}$ Extracted by organic solvents & modified

Identified by HPLC - Phenylthiohydantoin amino acid derivative

↓
sequence

Identified of N-terminal.

Maximum Amino acid Identification $\rightarrow 30$.

① Isolate the protein

↓
Purify it

↓
Proteins (3-D)

Breaking of disulfide bridge. $\leftarrow \rightleftharpoons \rightarrow$ β -mercapto ethanol treatment

Poly peptide
(Linear structure)

↓ Trypsin / CNBr treatment
Small peptides

↓
SDS page

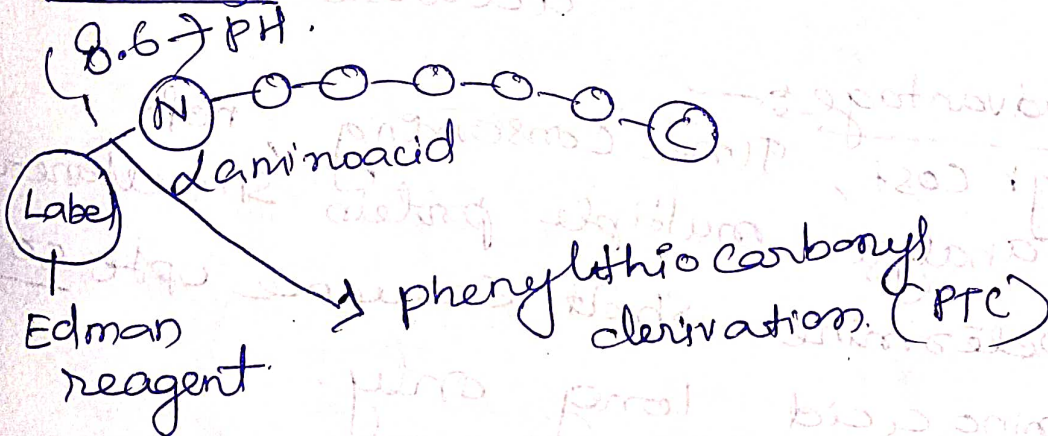
peptide fragments are obtained

Edman sequencing

— 3 steps — chemical method sequencing

① coupling ② cleavage ③ conversion

① coupling



② cleavage

Trifluoroacetic acid

Anhydrous — cut the condition N-terminus peptide bond.

↓
(AZT) Anilinothiazolinone

→ After separation of AZT derivative the remaining peptide is used for another round of reaction with (PITC)

(B) Conversion

→ AZT residue is separated from peptide by extraction in organic solvent (Ethyl acetate)

→ Then AZT residue is converted into more stable form phenylthiohydantoin (PTH derivative)

↓
Separated and Identified by HPLC
(High performance liquid chromatography)

Advantage

→ Amino acid - Information, sequence accurate

Disadvantages

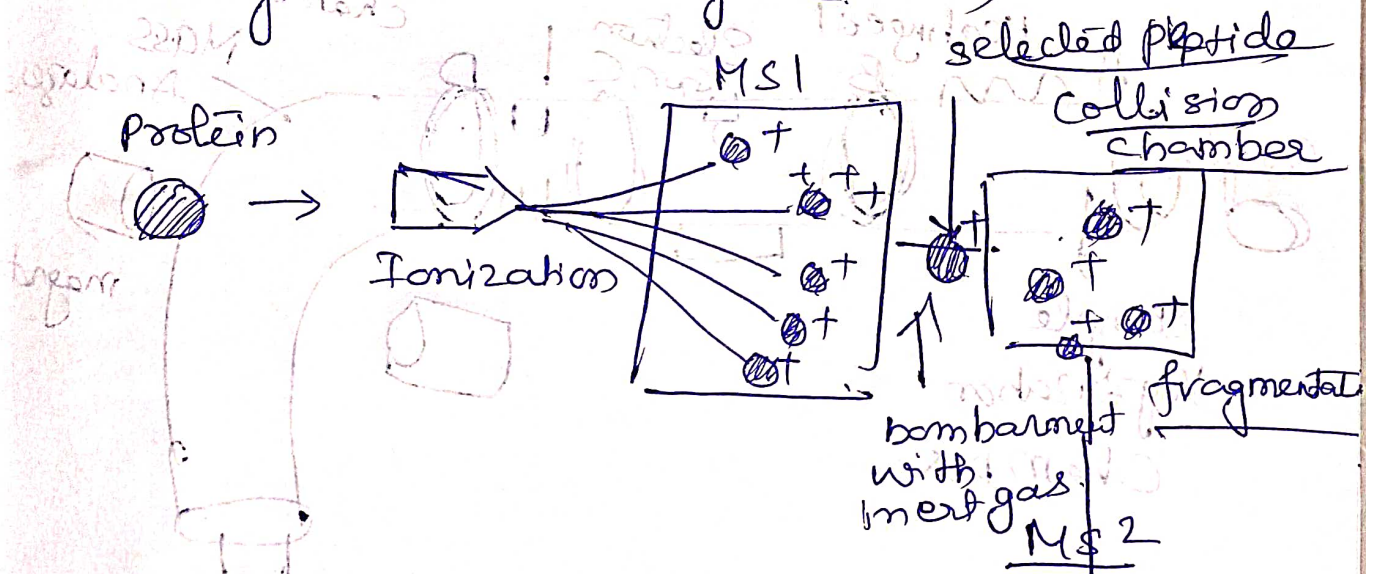
→ High cost, Time consuming, not analyze multiple protein simultaneously
→ Determines short sequence upto 50 amino acid long only.

Mass Spectroscopy — Physical method

- ① production of ions, in gaseous state
- ② Acceleration of Ions in magnetic & electric field
- ③ separation of ions in mass analyzer
- ④ Detection of Ions.

Process

- use to determine primary structure
- molecule sprayed through capillary tube
- forms gas-phase ions
- first Instrument sorts ions so that only one emerges (MS1)

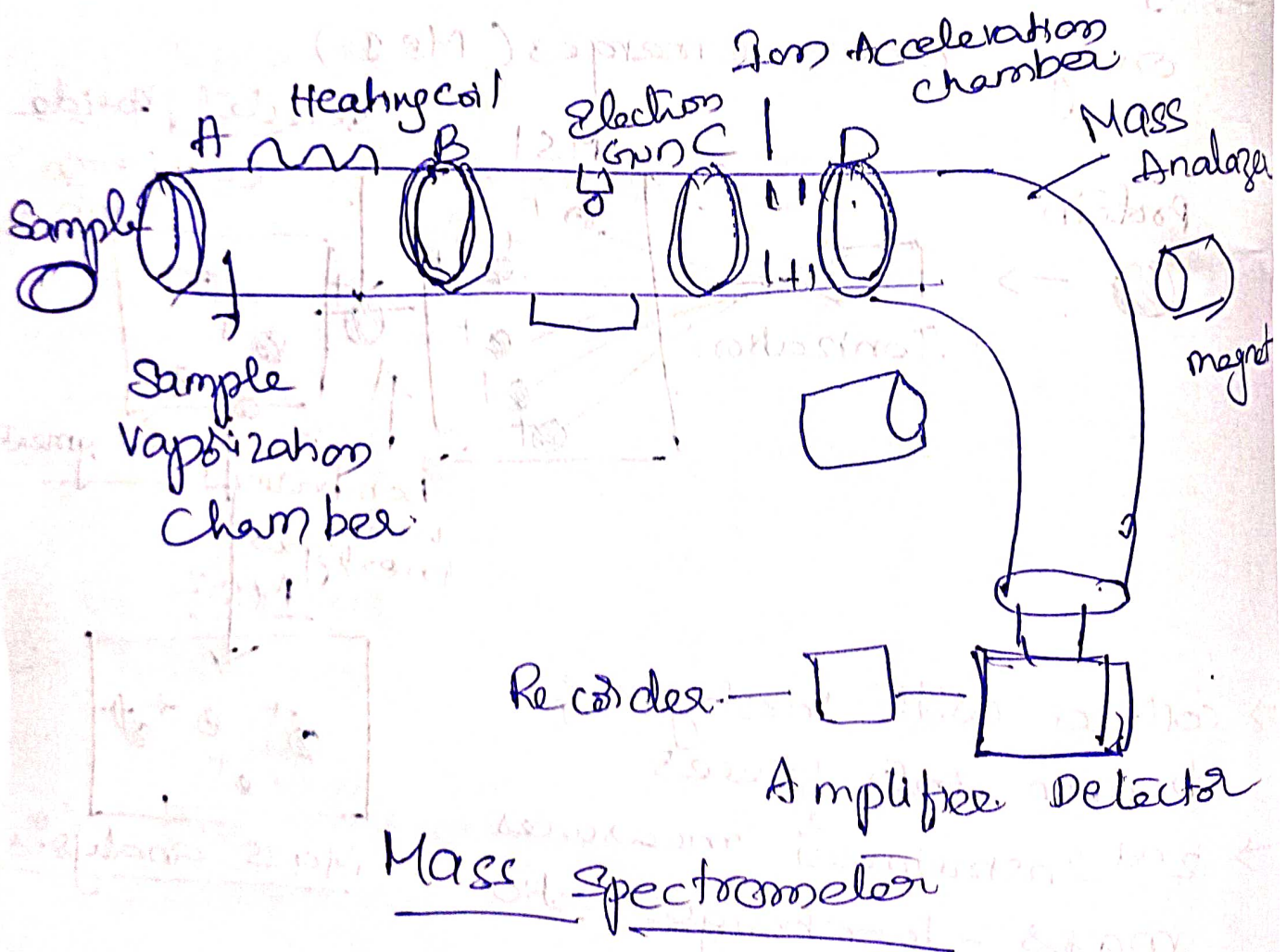
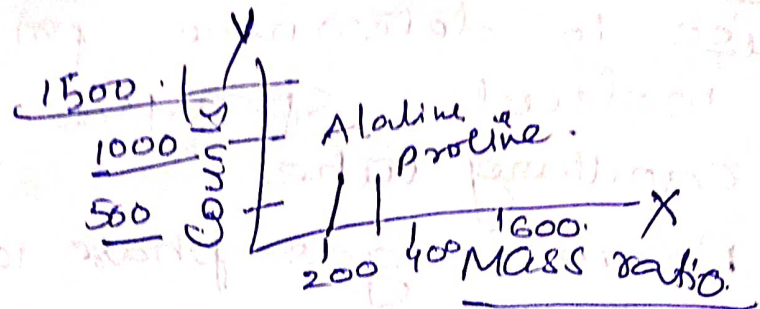


- collides with inert gas & broken into pieces

- 2nd Instrument measures mass-to-charge ratio of pieces (MS2)

- physical breakage of peptide
- can detect any modification.

- Measure charge to mass ratio
- sequence determined by comparing masses of increasingly larger fragments
- Difference in mass of successive peaks identifies each residue.



Protein annotation

— Protein annotations.

— computing protein sequences in the most common way to group and organize protein structures into families.