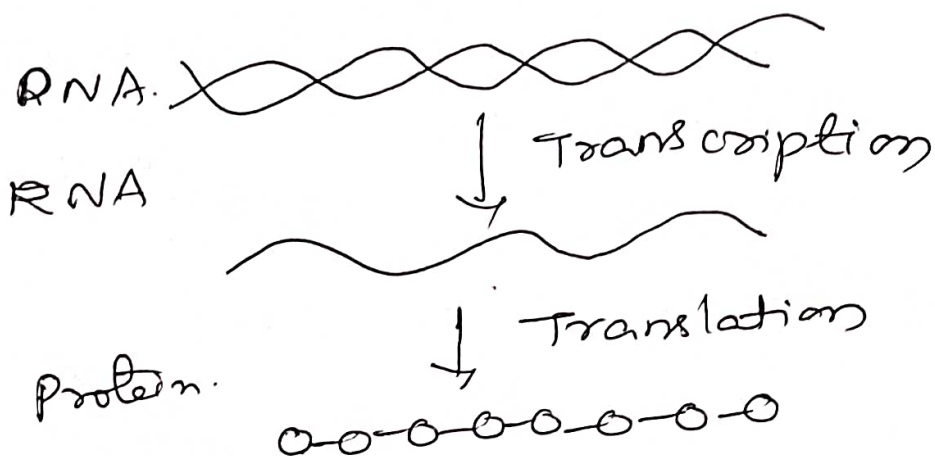


What is Sequence?

→ 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

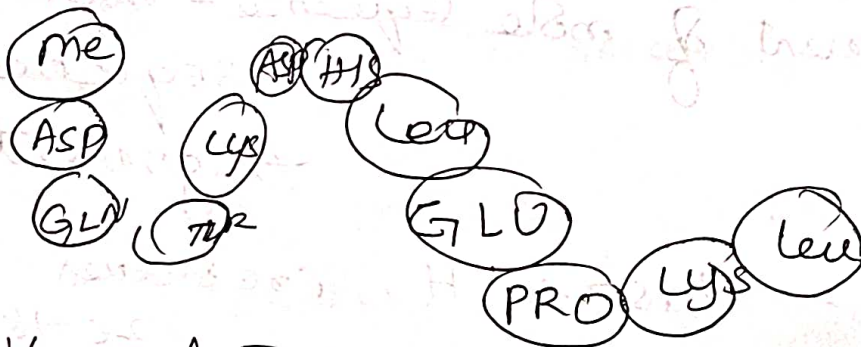
② for protein sequences:-

Both words have different meanings

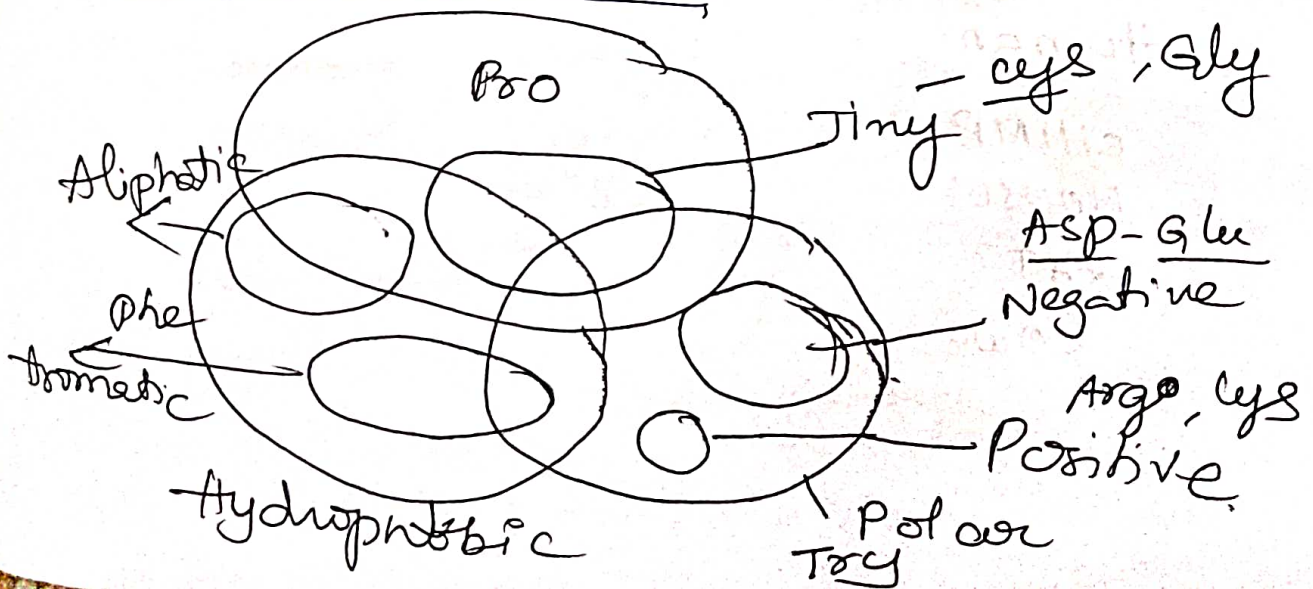
Identity :- % of exact matches b/w 2 aligned sequences

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
Identity

Strand A $\rightarrow$ A A G G C T T <sup>distance
mean</sup>

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% <sup>show
ident</sup>)

- 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

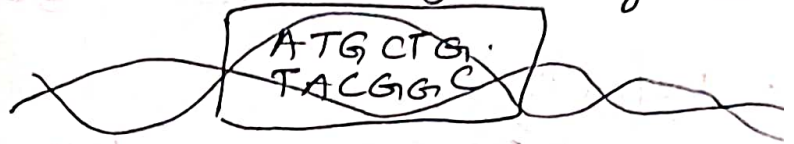
MOUSE

RAT

Cow.

why sequence alignment

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



→ function prediction

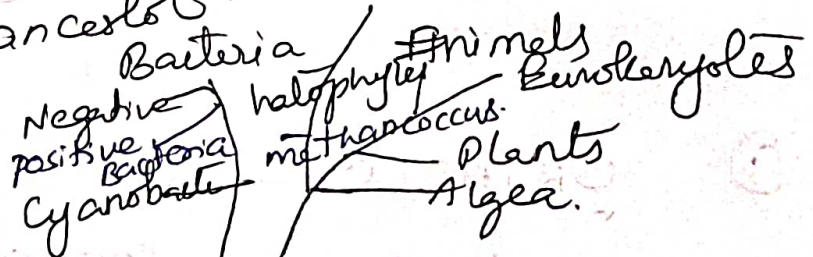
- gene - eg: Protein - functions.
enzyme
eg: Insulin, haemoglobin, RUBiSc,
- actin & myosin [contraction]
- ~~storage~~ ferritin (storage)
- immunoglobulin (protection)
Antibiotics.

→ 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.



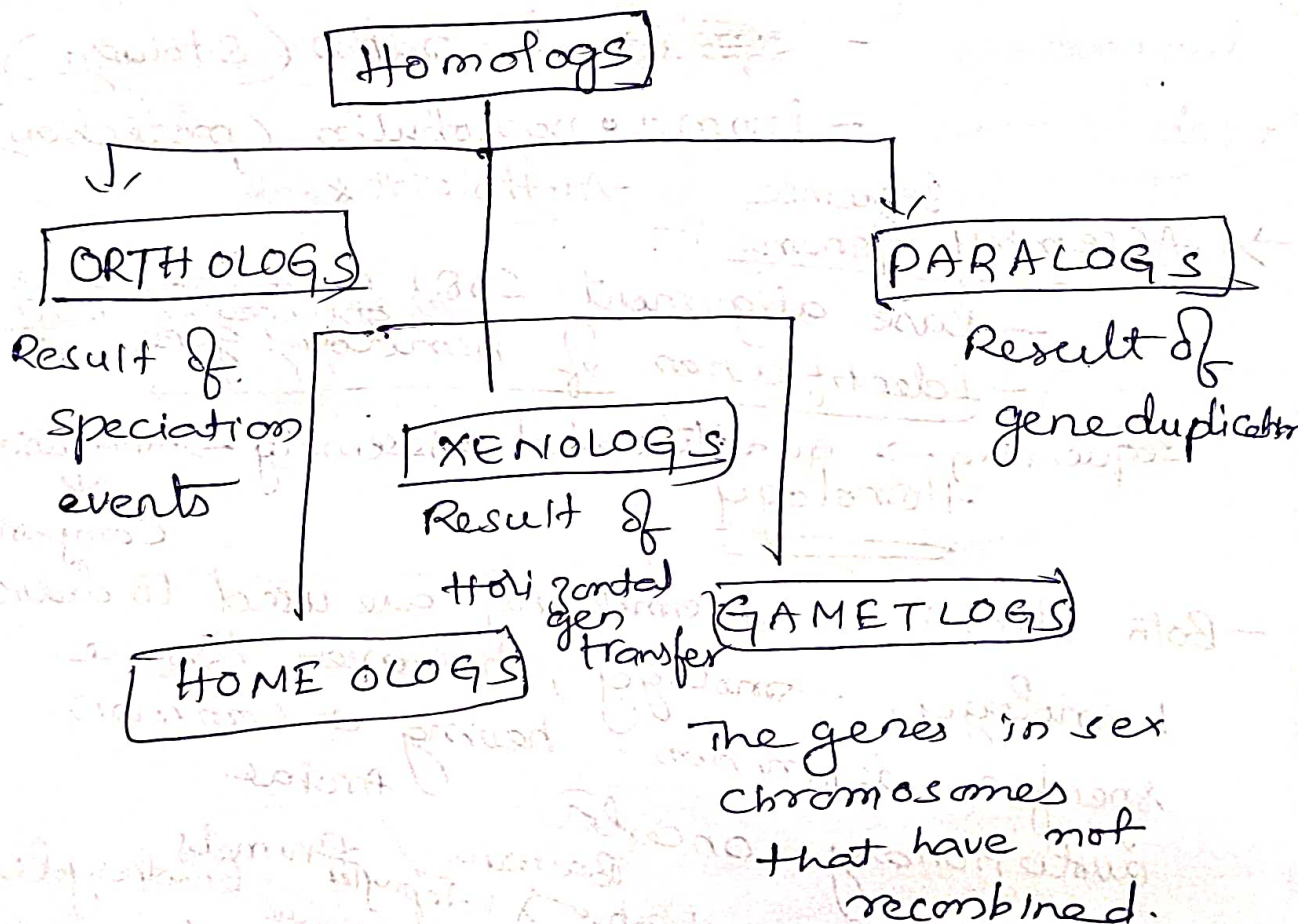
last common
Ancestor

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

Homology

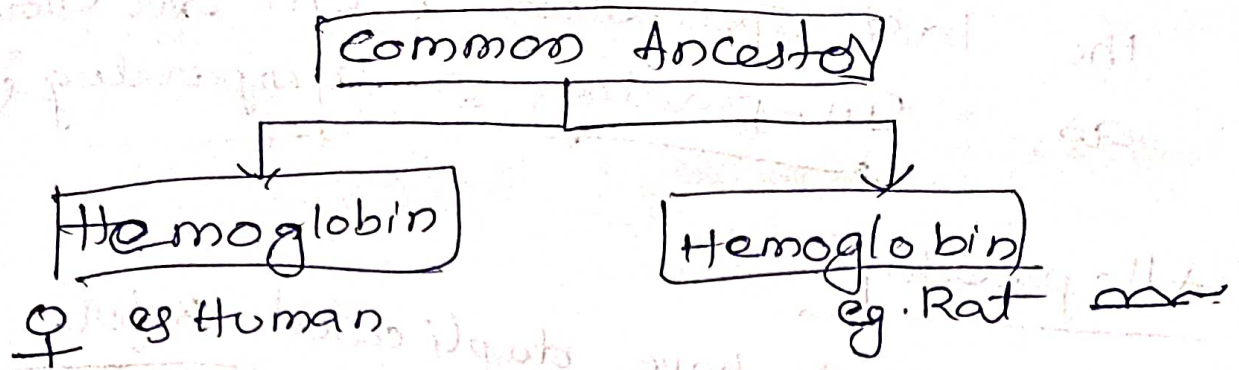
- Definition :-

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



① ORTHOLOGS :-

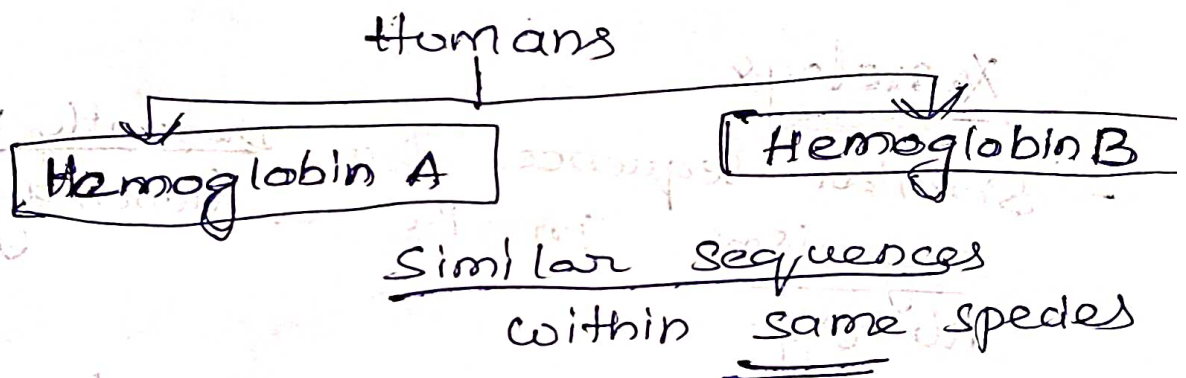
- Orthologous sequences are those sequences which are the results of the speciation process.



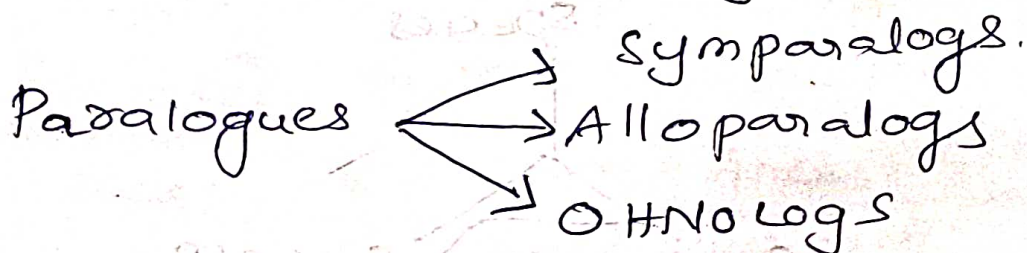
- Similar or identical sequences useful in different species
- construction of phylogenetic trees, information of species, evolution of genomes

PARALOGUES

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

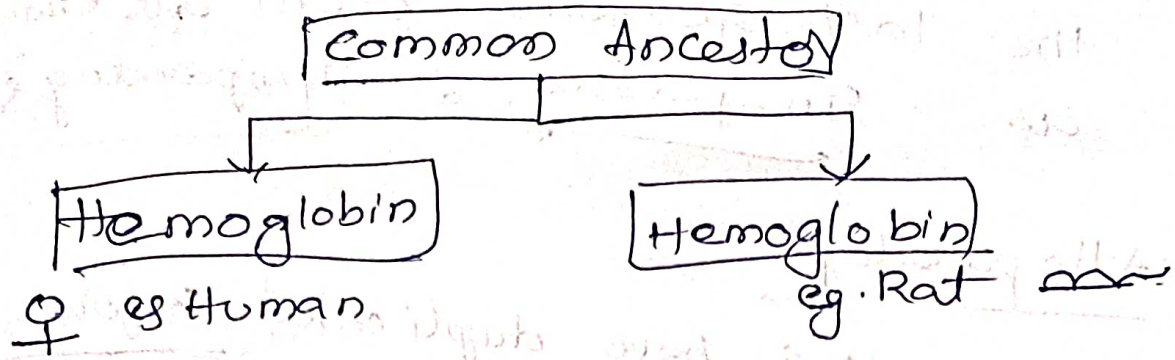


- protein evolution — Imp
- eg: — Hemoglobin A, B, A₂, B, F
- paralogous genes.



Symparalogs

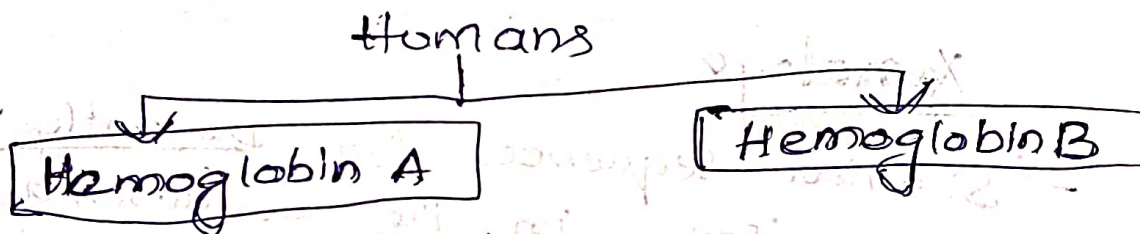
- Genes that have duplicated After



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

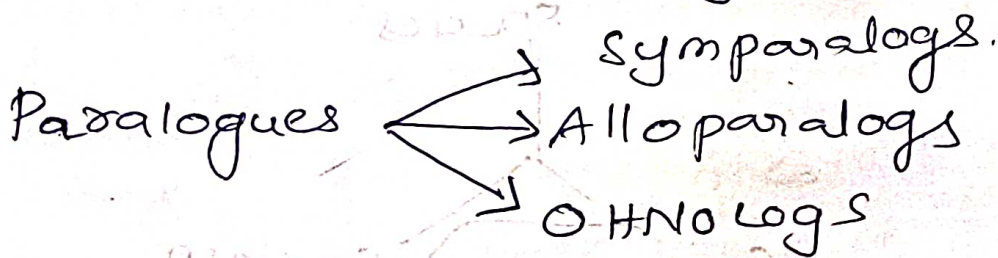
PARALOGUES

- 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
- paralogues 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.

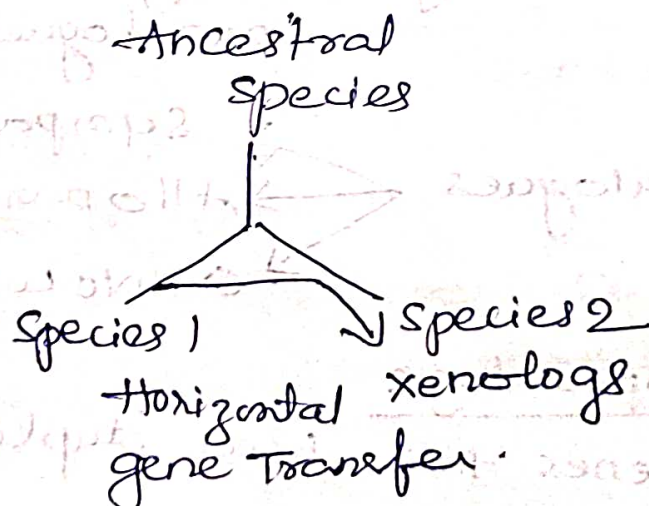
3. Ohnologs

→ 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

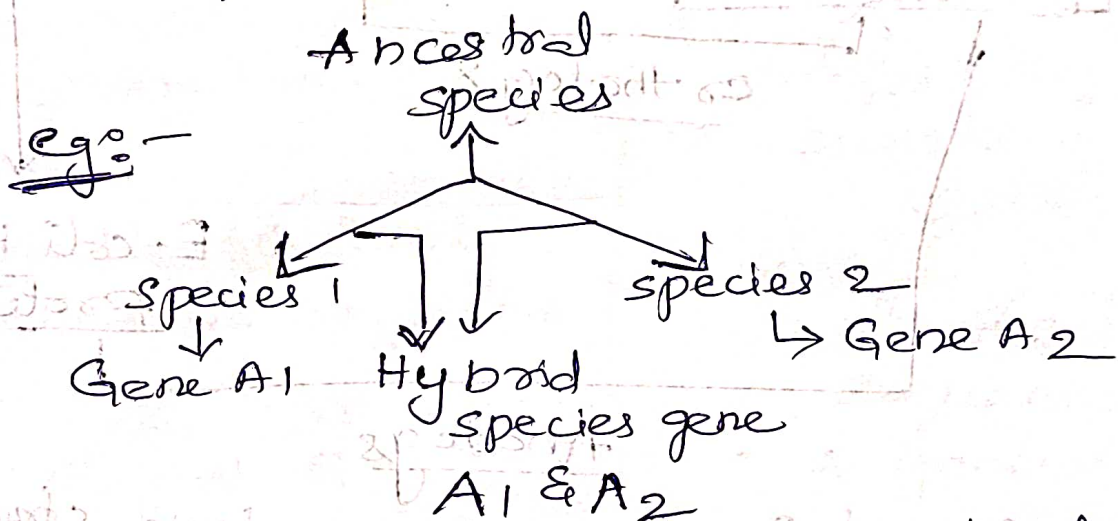


→ Lateral gene transfer → disease pathogens
 ↑
 Non-related organisms
 ↓
 Out breaks

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: ↓ Homologs.

Analogous :-

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

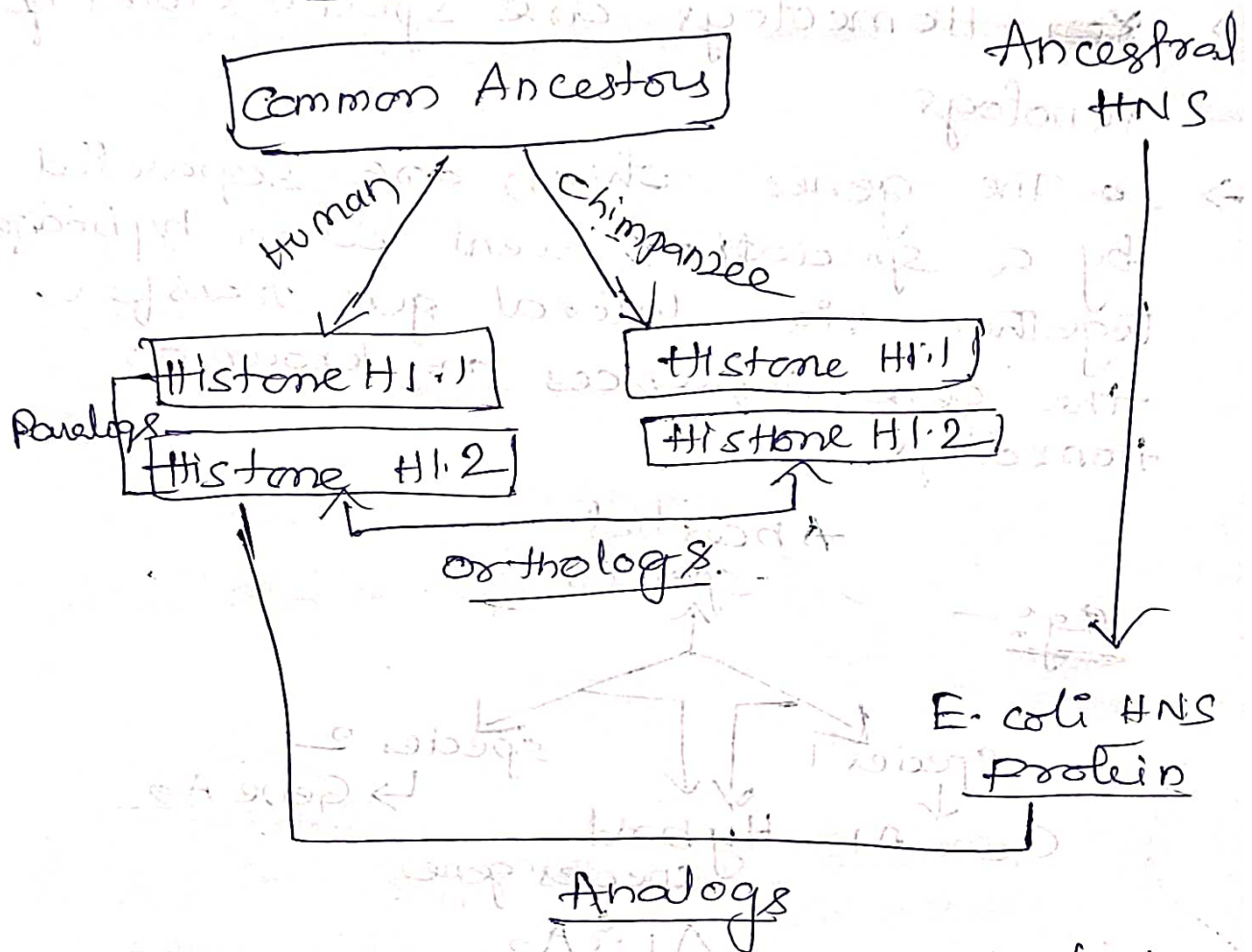
share a common Ancestor sequences.
 convergent Evolution

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 ① Pair-wise sequence alignment of two or more sequence.

→ 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  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

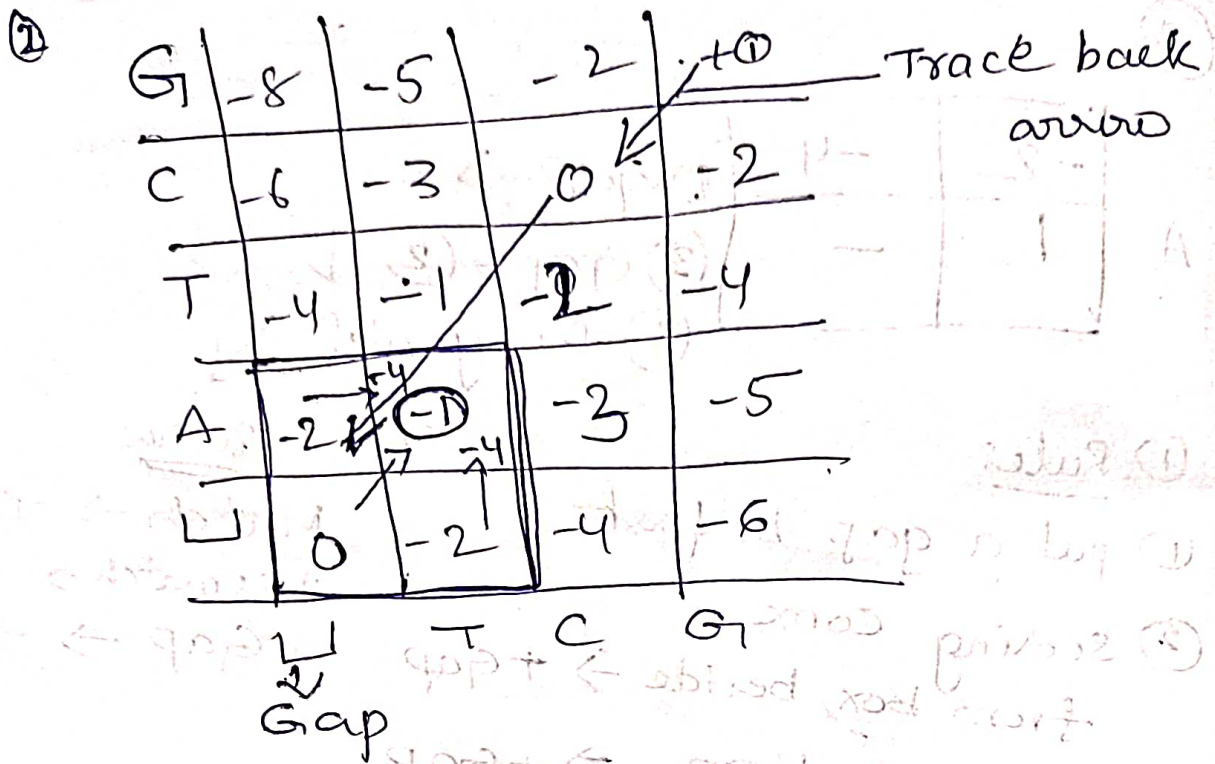
①

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

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

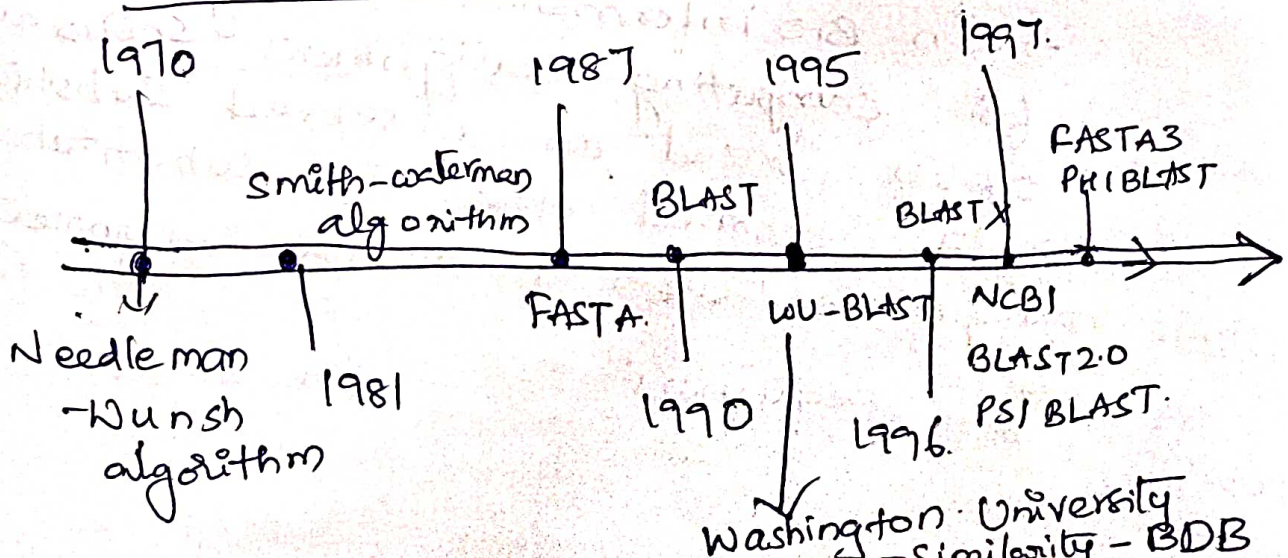
Match → 1

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

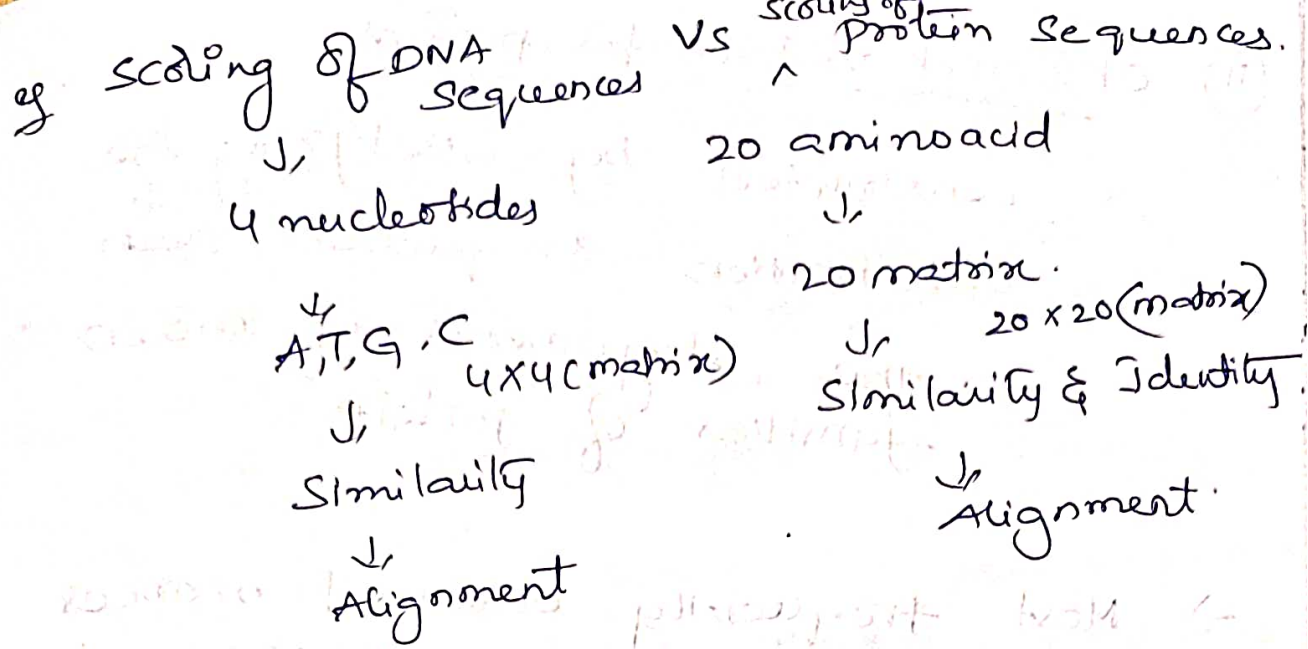
→ Sequence alignment to determine similarity between sequences.

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

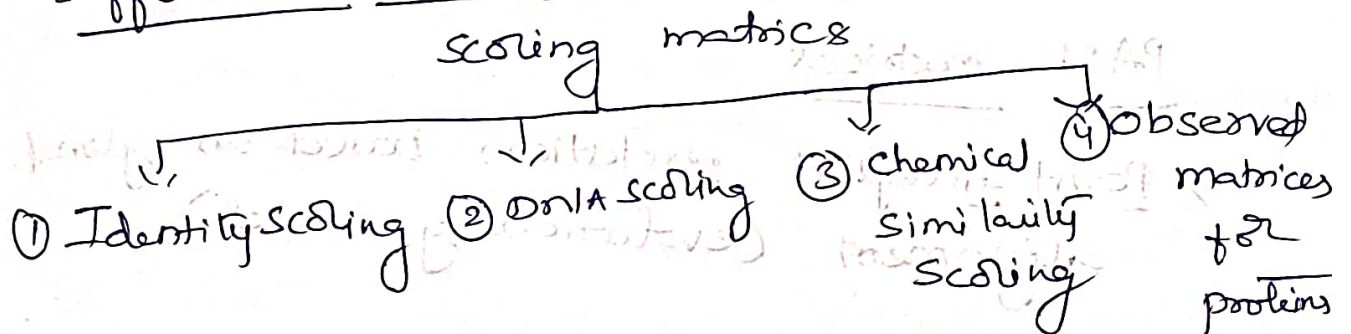
→ we must provide a score to each similar sequence and this score is provided using scoring matrix.

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 proteins

→ 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.

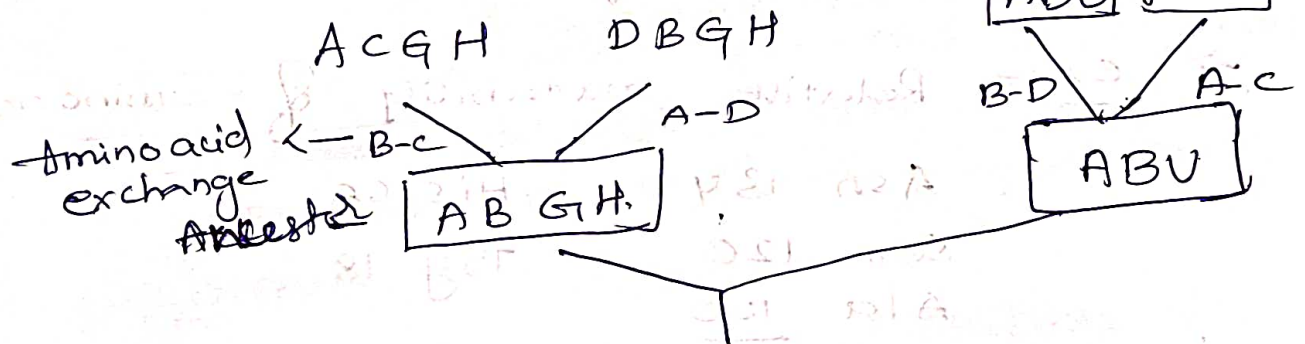
- construction of PAM matrix involves alignment of full length sequence & subsequent construction of phylogenetic trees using likelihood principle.
- PAM accepted by the natural selection.
- A PAM is defined as 1% amino acid change or one mutation per 100 residues.

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						
D		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 Trp 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.

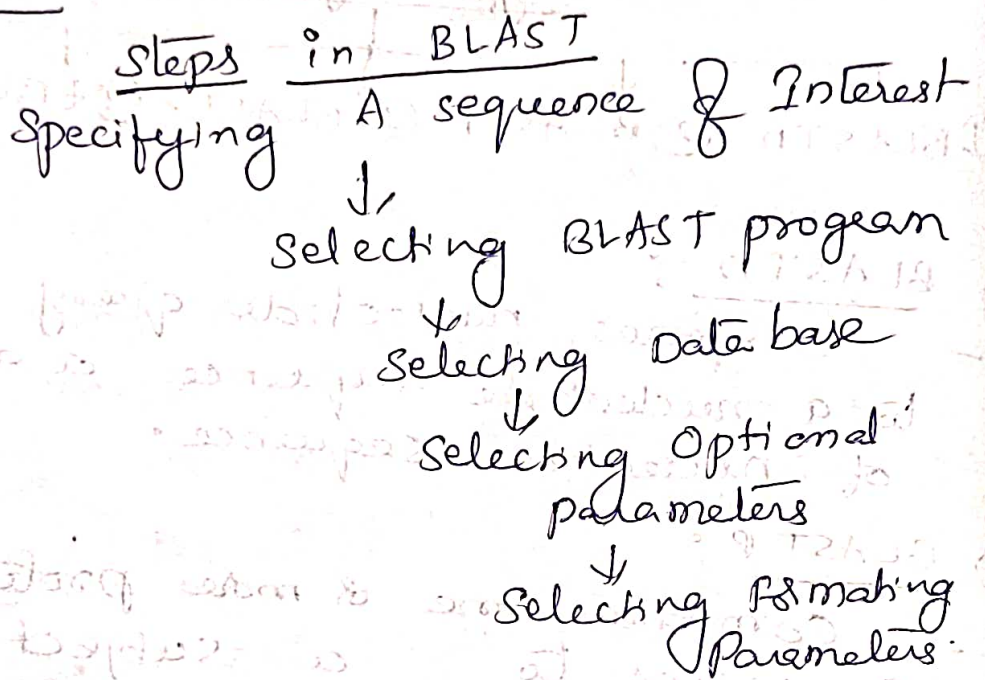
④ tBLASTn :

- 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.