أكاديمية كاوست
KAUST ACADEMY

Introduction to Bioinformatics

Day 2: Sequence Alignment and Phylogenetic

13th January 2026

KAUST

Outline Day 2

Morning (9-12 pm):

Pairwise and Multiple Sequence Alignment:

- **Overview of alignment concepts:** Pairwise alignment (global vs local) and multiple sequence alignment
- **Algorithm**
  - **Global Alignments**: Needleman-Wunsch algorithm
  - **Local Alignments**: Smith-Waterman algorithm
  - **Multiple Sequence Alignment**: ClustalW and Clustal Omega

Practical Session:

Primary option:

- Use NCBI BLAST (web-based) to perform local alignment and sequence similarity searches
- Perform multiple sequence alignment using Clustal Omega (web-based)
- Code with python

Afternoon (2-5 pm): Phylogentics

Phylogentics:

- Introduction to Phylogentics tree concept, and significance in evolution biology.

Practical Session:

- Construct phylogentics trees using Clustal Omega (web-based) starting from multiple sequence alignments.
- Code with python

أكاديمية كاوست
KAUST ACADEMY

Day 2: Sequence Alignment and Phylogenetic

Morning Session

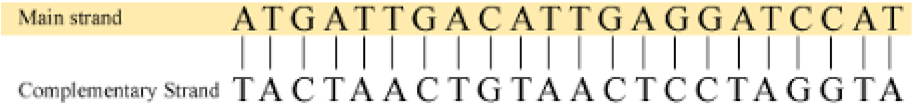
KAUST

What is a Sequence?

Specific linear order of nucleotides (in DNA or RNA) or amino acids (in proteins) that determines the structure and function of these biomolecules.

DNA Sequence: The arrangement of bases (adeniine [A], thymine [T] cytosine [C], guanine [G]).
RNA Sequence: The arrangement of bases (adenine [A], uracil [U], cytosine [C], guanine [G]).
Protein Sequence: The arrangement of amino acids in a polypeptide chain.

Sequences encode genetic information and are fundamental to biological processes like replication, transcription, and translation.


Main strand ATGATTGACATTGAGGATCCAT
Complementary Strand TACTAACTGTAACTCCTAGGTA

Sample genetic code with complementary strands.

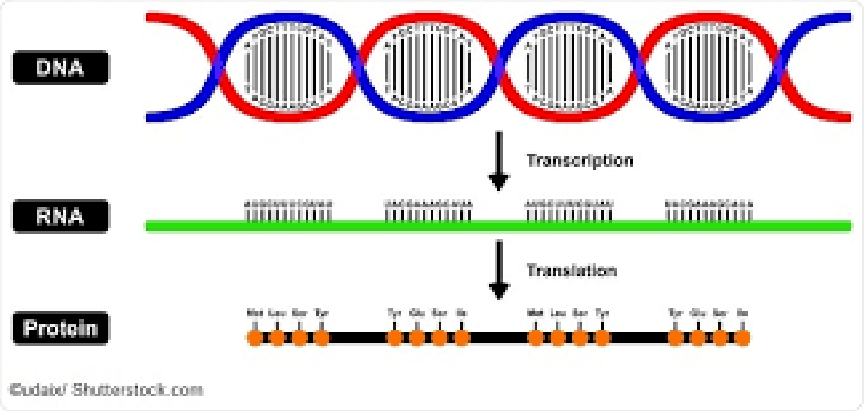
© G.Osuri

Examples of Sequences?

DNA Sequence: order of nucleotide bases (adenine, thymine, cytosine, guanine)
Example. ATCGTACGGA

RNA Sequence: Similar to, DNA1 but thymine (T) is replaced by uracil (U) 'in RNA.
Example: AUCGAUCGGA

Protein Sequence: order ,of amino acids (e.g., methionine, alanine, leucine,,glycine)
Example· MET-ALA-LEU-GLV



©udaix/ Shutterstock.com

FASTA

A text-based file format used to represent nucleotide or protein sequences. It is commonly used in bioinformatics to store and exchange sequence data. A FASTA file consists of:

1. Header Line: Starts with a > symbol, followed by an identifier or description of the sequence.

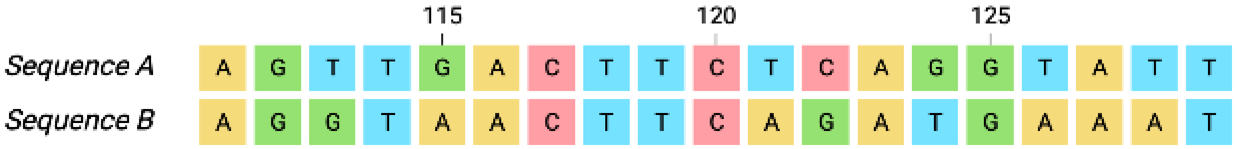
2. Sequence Data: The following lines contain the actual sequence, either nucleotide (DNA/ RNA) or protein, and can span multiple lines.

Example:

```
>Sequence_ID_1 Description of the first sequence AGCTAGCTAGCTAGCTAGCTAGCTAGCTAGCTA
>Sequence_ID_2 Description of the second sequence TGCATGCATGCATGCATGCGTACGATCGTACG
```

What is Sequence Alignment?

Sequence alignment is the process of arranging two or more biological sequences such as nucleotide (DNA, RNA), or amino acid (protein) sequences to identify regions of similarity.



					115					120					125				
Sequence A	A	G	T	T	G	A	C	T	T	C	T	C	A	G	G	T	A	T	T
Sequence B	A	G	G	T	A	A	C	T	T	C	A	G	A	T	G	A	A	A	T

What alignments can help?

DETERMINE EVOLUTIONARY RELATIONSHIPS AMONG GENES, PROTEINS, AND SPECIES

DETERMINE FUNCTION OF A NEWLY DISCOVERED GENE SEQUENCE

PREDICTING STRUCTURE AND FUNCTION OF PROTEIN

Importance of Sequence Alignment in Bioinformatics

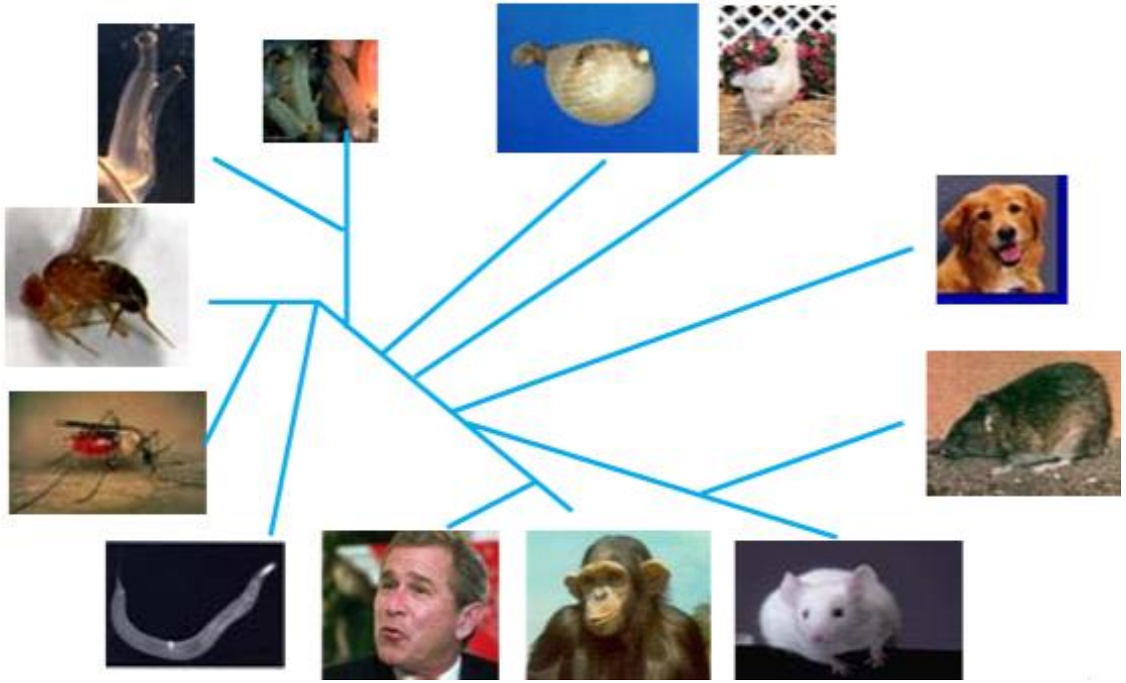
Evolutionary Relationships

- Helps in identifying similarities between genes/
- proteins from different species.

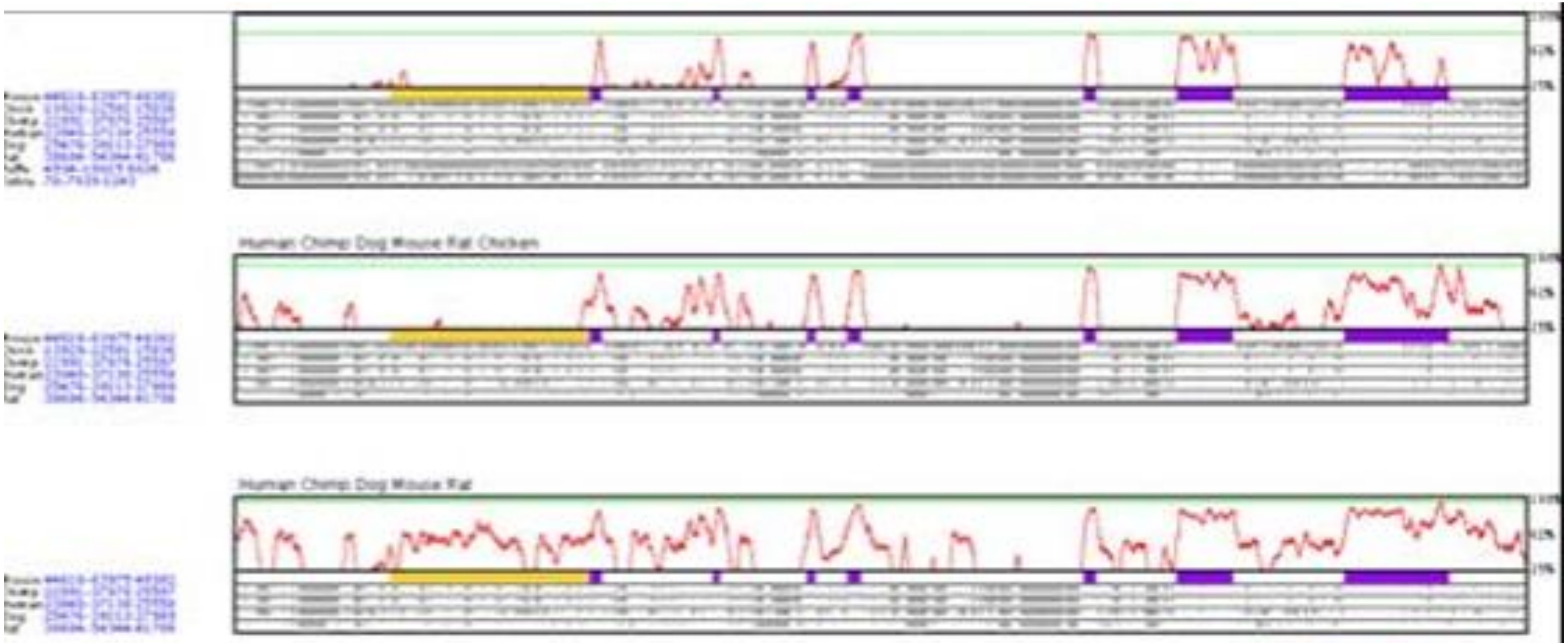
- Homologous sequences,(genes from a common ancestor) can be identified through alignment.

- Helps inferring evolutionary distances using substitution patterns (changes in nucleotides).

- Helps in phylogenetic tree construction

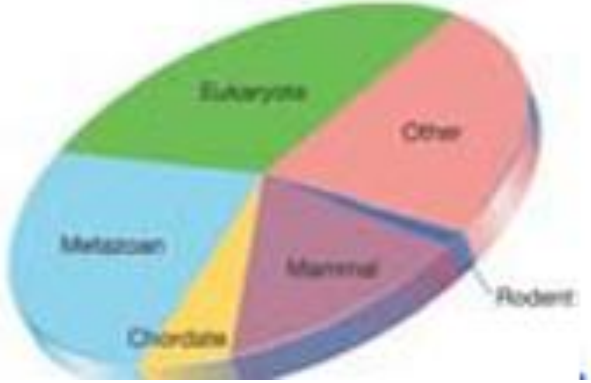


Sequence conversation implies function



Alignment is the key to

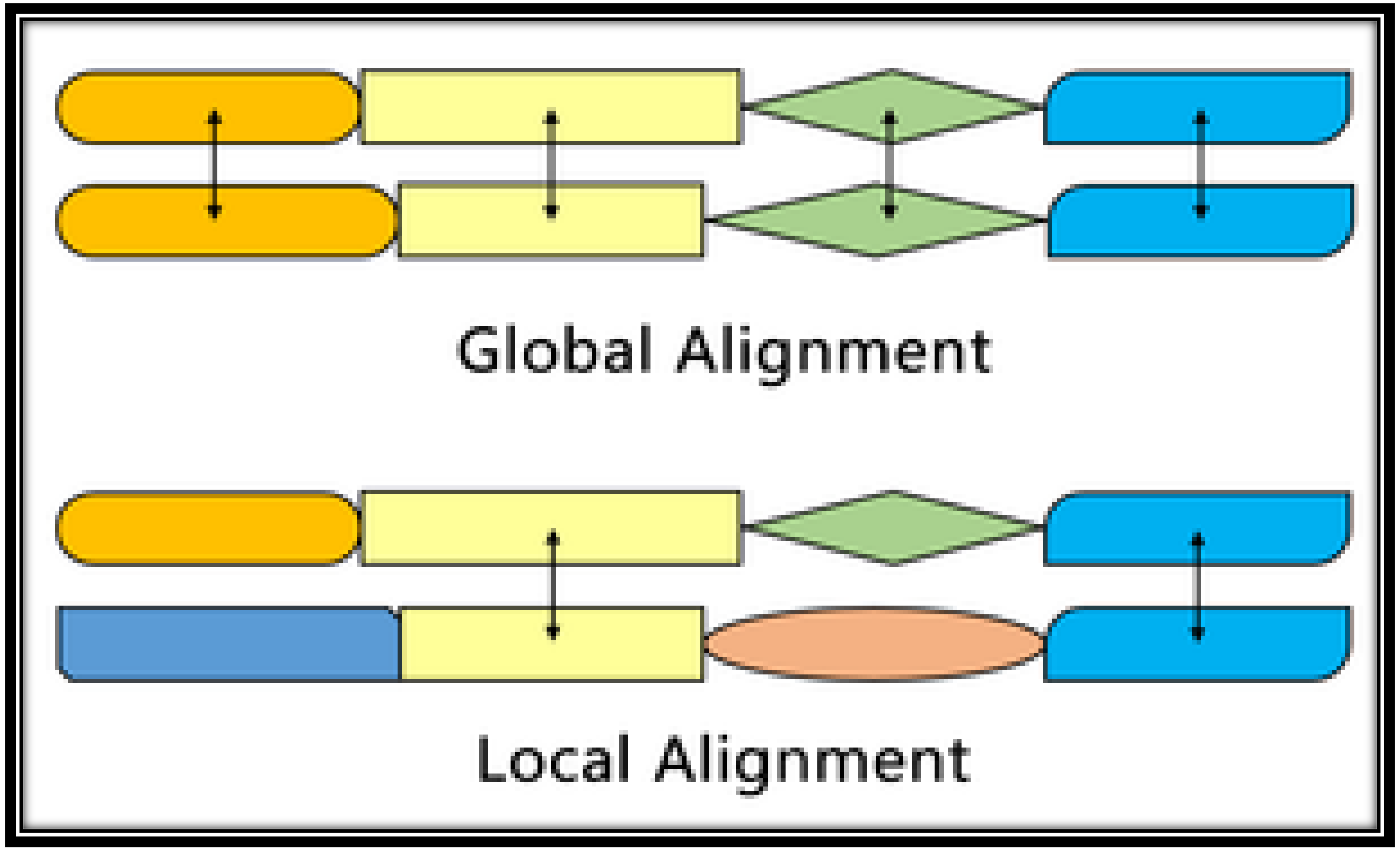
- Finding important regions
- Determining function
- Uncovering the evolutionary forces



Algorithm used in Dynamic programming

Needleman-wunsch Algorithm	Smith-Waterman algorithm
Developed by Saul.b. needleman & christian.d. wunsch	Developed by Temple.F.Smith & Michael. S. Waterman
Referred as global alignment	Referred as local alignment
Used in aligning 2 closely related sequence	Used in aligning divergent sequences
Compares the whole sequence	Compares a patch from the sequence
Tools:- EMBOSS-Needle, Specialised BLAST	Tools:-EMBOSS-Water, LALIGN

Comparison of Global and Local alignment in general



Needleman-Wunsch algorithm

- Aligns protein or nucleic acid sequences
- Divides a large problem into a series of smaller problems
- Uses the solution of smaller problem to reconstruct a solution to the larger problems

Constructing the matrix

We will have 2 matrices of 2D representation viz,

1. The score matrix
2. Traceback matrix

The N-W algorithm consists of 4 steps:-

1. Initialization of the score matrix
2. Filling up the matrix
3. Traceback
4. Alignment

1. Initializing the scoring matrix

		G	C	A	T	G	C
	0	-1	-2	-3	-4	-5	-6
G	-1						
T	-2						
A	-3						
C	-4						
G	-5						
C	-6						

Match = 1
Mismatch. = -1
Gap. = -1

2. Filling the matrix

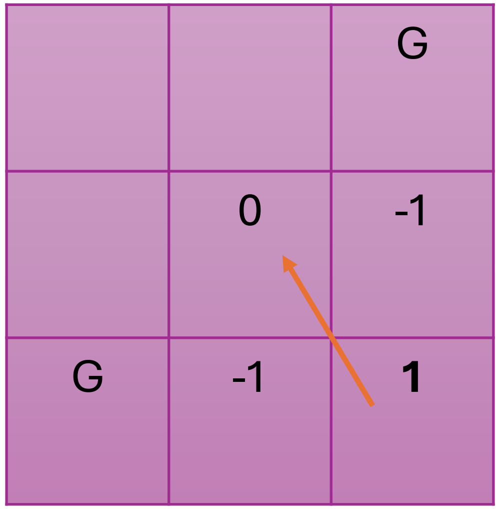
		G
	0	-1
G	-1	X

For x :

This cell has 3 possible values

- Top :- (-1)+(-1) = -2
- Left :- (-1)+(-1) = -2
- Top-left :- (0)+(1) = 1

The highest value is 1 and thus it is entered into the cell i.e x = 1



		G
	0	-1
G	-1	**1**

Contd..

		G	C
	0	-1	-2
G	-1	1	X
C	-2	Y	

For **X** :

Top: (-2)+(-1) = (-3)

Left: (+1)+(-1) = (0)

Top-Left: (-1)+(-1) = (-2)

For **Y** :

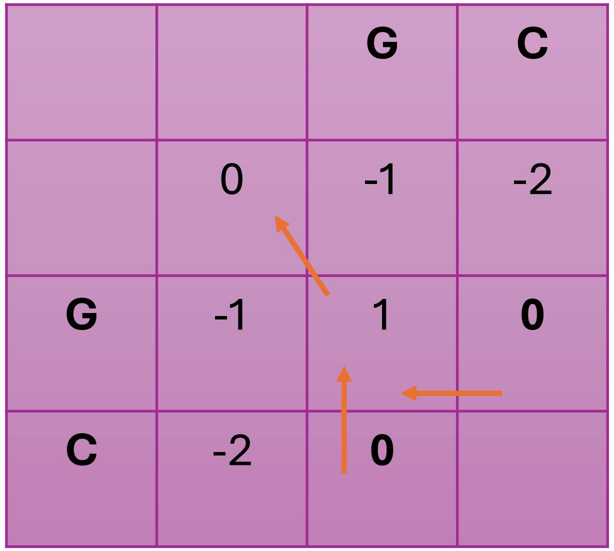
Top: (1)+(-1) = (**0**)

Left: (-2)+(-1) = (-3)

Top-Left: (-1)+(-1) = (-2)

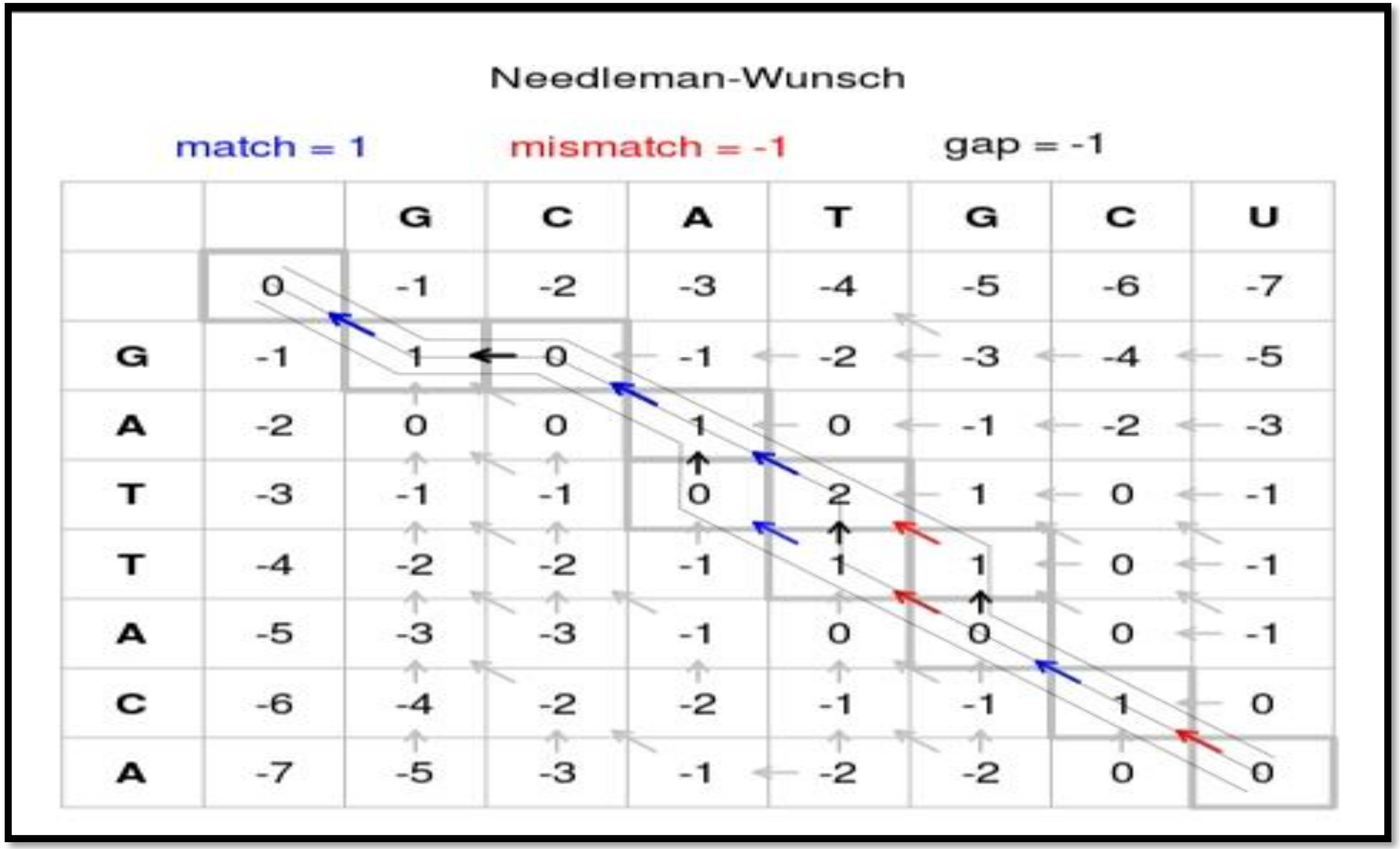
The highest value for X and Y is 0, thus it is entered into the cell.

i.e **X = O; Y = 0**



		G	C
	0	-1	-2
G	-1	1	**0**
C	-2	**0**	

3. Traceback



Needleman-Wunsch

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

4. Alignment

Rules :-

1. If arrow is vertical/horizontal assign a gap and a character

↑ / → :- Gaps and characters

Where to Assign

a gap and a character ?

Sequenc
e 1 or 2

Contd…

Ans) The gap will be assigned in the direction of the arrow and the character will be assigned in the opposite direction

2. If there is a diagonal arrow both the characters will be assigned

:- Both characters

↙

Result of alignment

```
G C T A G C -
      | . | |
          .
G - T A C G C
```

Smith-Waterman algorithm

- S-W algorithm is modified version of Needleman-Wunsch algorithm
- Negative scoring matrix cells are set to zero
- Traceback procedure starts at the highest scoring matrix cells and procedure until a cell with score zero is found

Constructing the matrix

		G	C	A	T	G	C
	0	0	0	0	0	0	0
G	0						
C	0						
A	0						
T	0						
G	0						
C	0						

Match	= 1
Mismatch.	= -1
Gap.	= -1

Contd..

		G	C
	0	-1	-2
G	-1	1	X
C	-2	Y	

For **X :**

Top: (-2)+(-1) = (-3)

Left: (+1)+(-1) = (0)

Top-Left: (-1)+(-1) = (-2)

For **Y :**

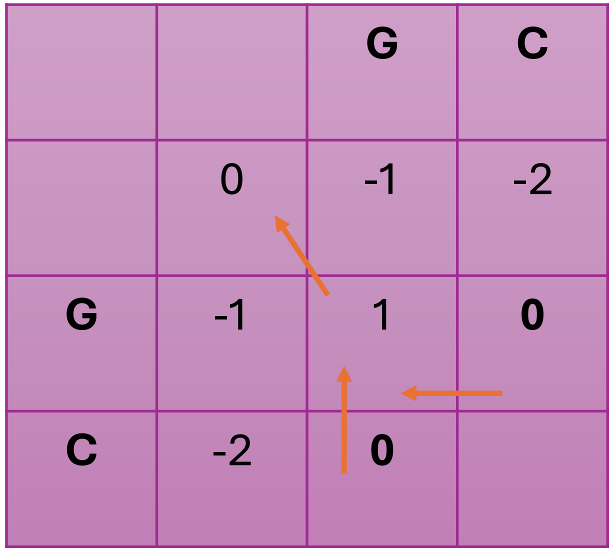
Top: (1)+(-1) = (**0**)

Left: (-2)+(-1) = (-3)

Top-Left: (-1)+(-1) = (-2)

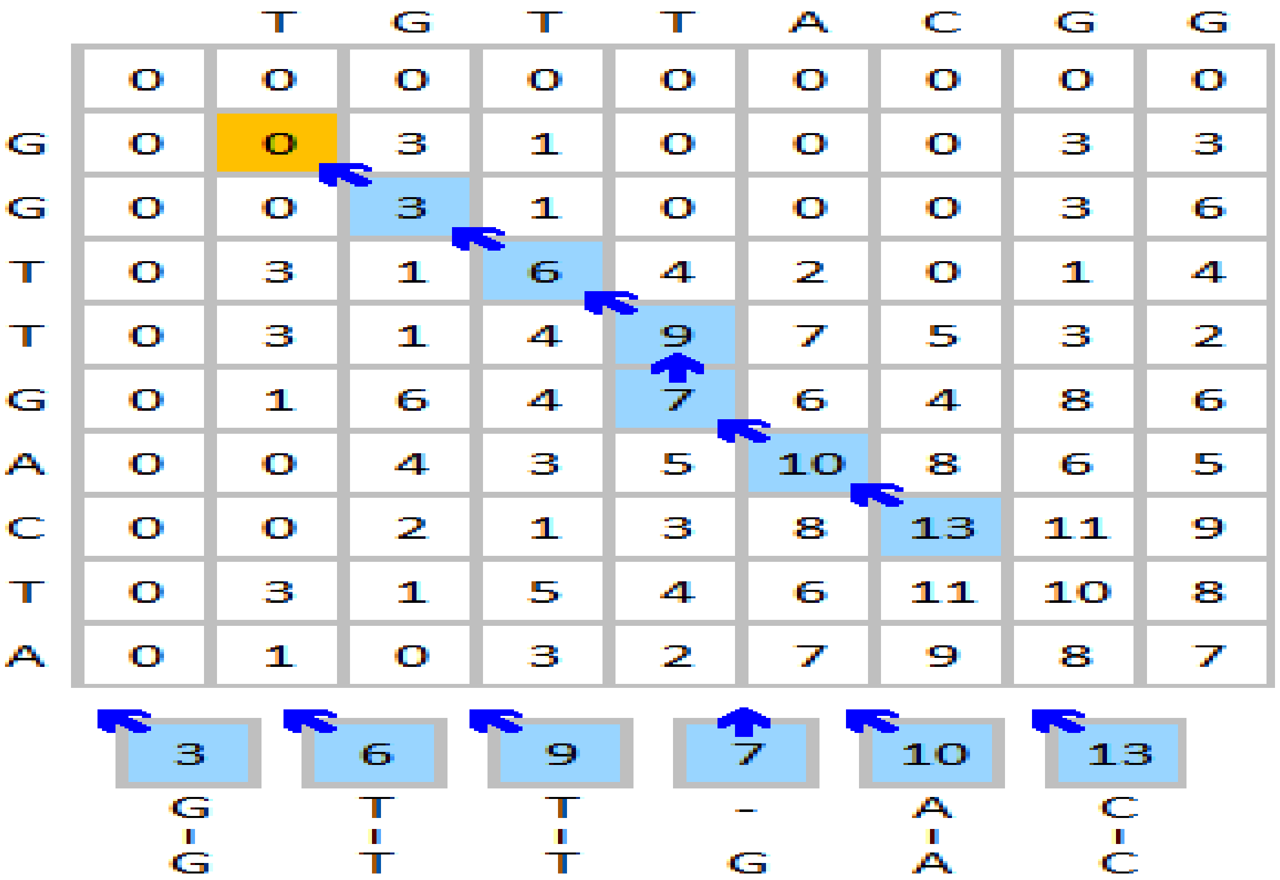
The highest value for X and Y is 0, thus it is entered into the cell.

i.e **X = O; Y = 0**



		G	C
	0	-1	-2
G	-1	1	**0**
C	-2	**0**	

Traceback



		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

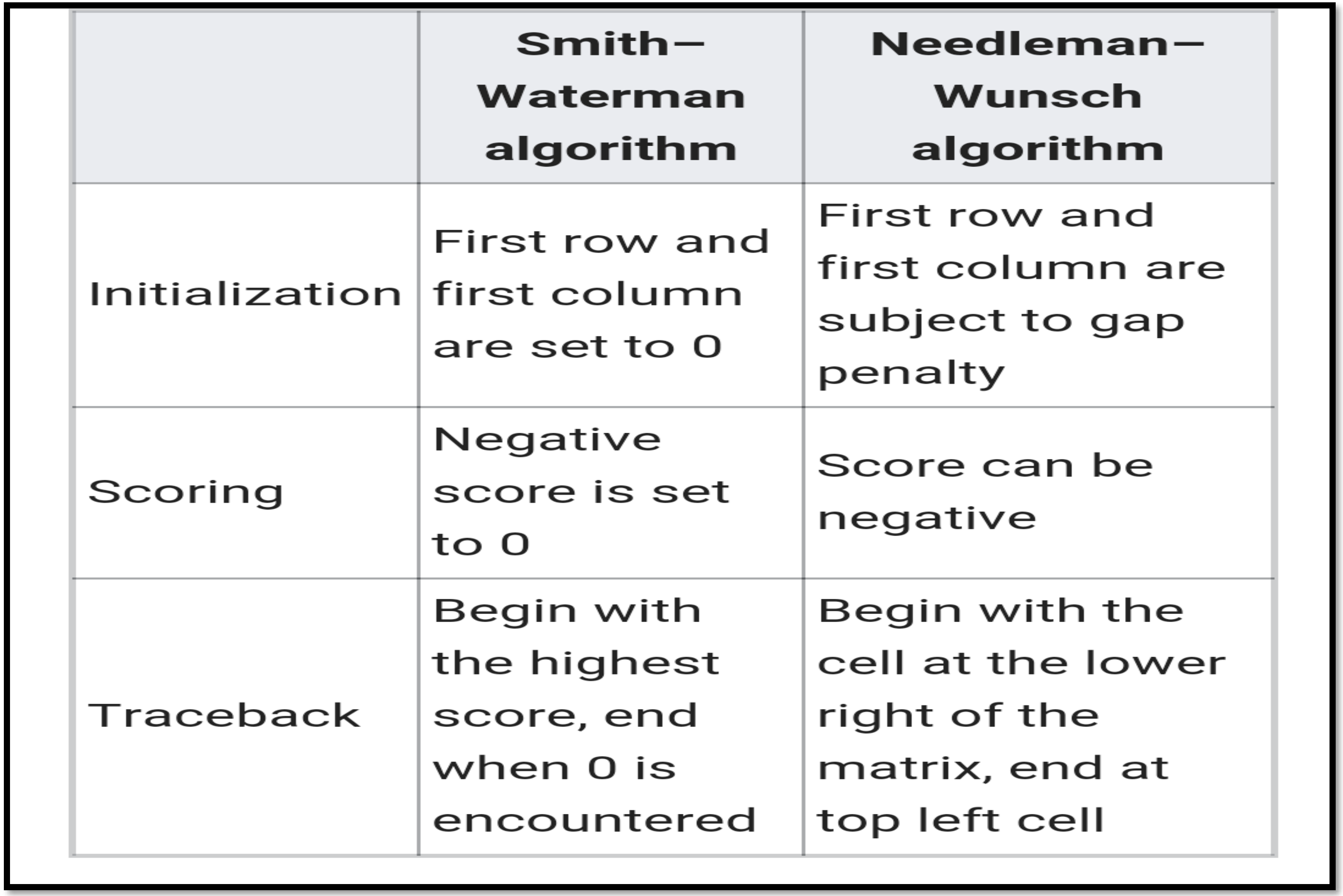
3	6	9	7	10	13
G	T	T	-	A	C
|	|	|		|	|
G	T	T	G	A	C

Alignment

Rules:

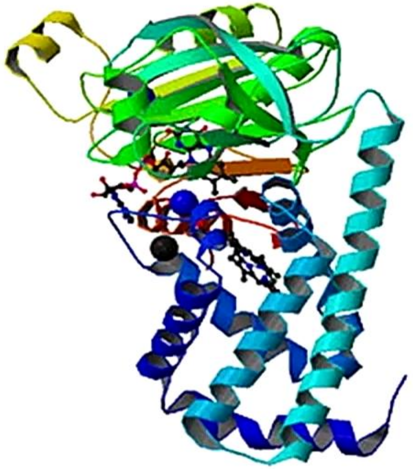
Same as of Needleman- Wunsch algorithm

Difference between the procedure of S-W and N-W algorithm

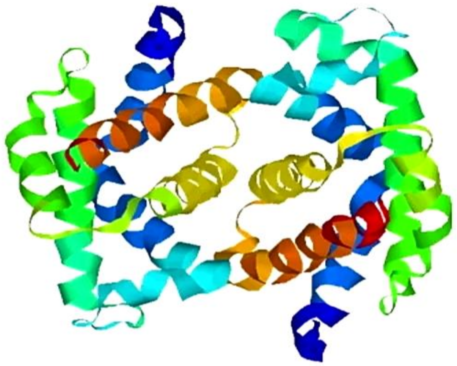


	Smith–Waterman algorithm	Needleman–Wunsch algorithm
Initialization	First row and first column are set to 0	First row and first column are subject to gap penalty
Scoring	Negative score is set to 0	Score can be negative
Traceback	Begin with the highest score, end when 0 is encountered	Begin with the cell at the lower right of the matrix, end at top left cell

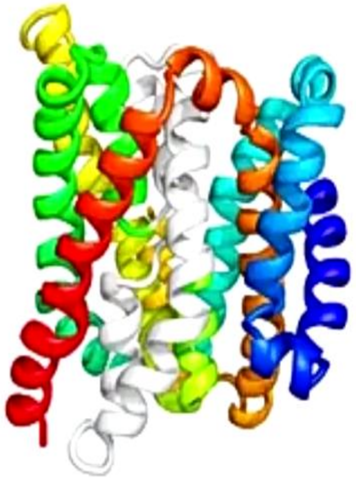
Example

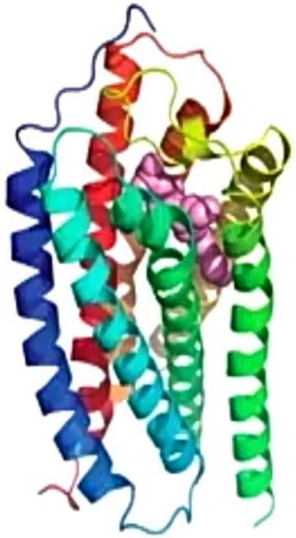


Flavohemoprotein of
Escherichia coli



haemoglobin subunit alpha
Homo sapiens

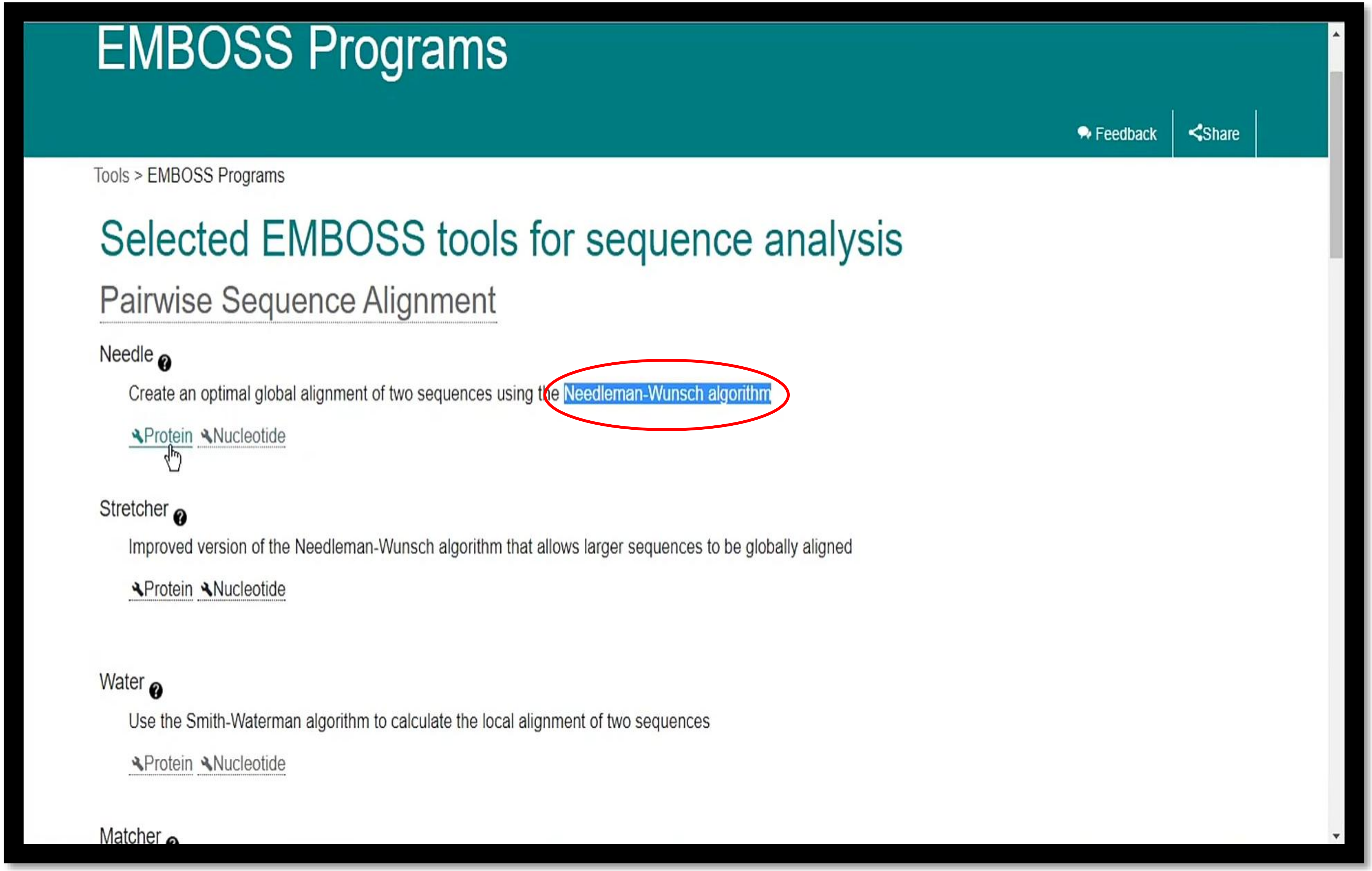




MNPYIYLGGAILAEVIGTTLMKFSEGF
TRLWPSVGTIICYCASFWLLAQTLAYIP
TGIAYAIWSGVGIVLISLLSWGFFGQRL
DLPAIIGMMLICAGVLIINLLSRSTPH

MSEALKILNNIRTLRAQARECTLETLEEMLE
KLEVVVNERREEESAAAAEVEERTRKLQQY
REMLIADGIDPNELLNSLAAVKSGTKAKRA
QRPAKYSYVDENGETKTWTGQGRTPAVIK

To align two sequences of 300 aa/nt lengths, approximately 10^{88} comparisons are to be made



EMBOSS Programs

Tools > EMBOSS Programs

Selected EMBOSS tools for sequence analysis

Pairwise Sequence Alignment

Needle

Create an optimal global alignment of two sequences using the Needleman-Wunsch algorithm

Stretcher

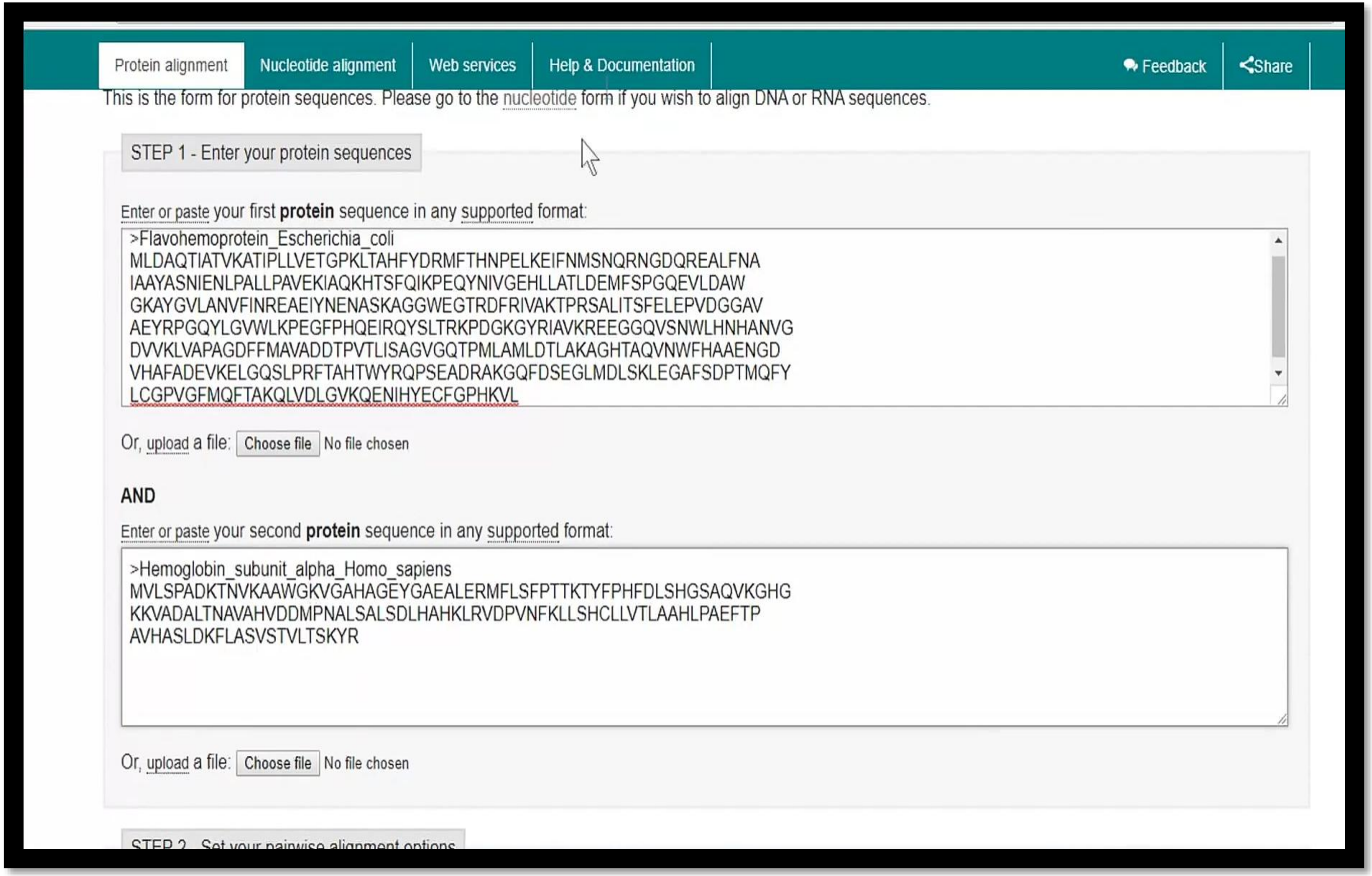
Improved version of the Needleman-Wunsch algorithm that allows larger sequences to be globally aligned

Water

Use the Smith-Waterman algorithm to calculate the local alignment of two sequences

Matcher

Emboss homepage



This is the form for protein sequences. Please go to the nucleotide form if you wish to align DNA or RNA sequences.

STEP 1 - Enter your protein sequences

Enter or paste your first **protein** sequence in any supported format:

```
>Flavohemoprotein_Escherichia_coli
MLDAQTIATVKATIPLLVETGPKLTAHFYDRMFTHNPELKEIFNMSNQRNGDQREALFNA
IAAYASNIENLPALLPAVEKIAQKHTSFQIKPEQYNIVGEHLLATLDEMFSPGQEVLDAW
GKAYGVLANVFINREAEIYNENASKAGGWEGTRDFRIVAKTPRSALITSFELEPVDGGAV
AEYRPGQYLGVWLKPEGFPHQEIRQYSLTRKPDGKGYRIAVKREEGGQVSNWLHNHANVG
DVVKLVAPAGDFFMAVADDTPVTLISAGVGQTPMLAMLDTLAKAGHTAQVNWFHAAENGD
VHAFADEVKELGQSLPRFTAHTWYRQPSEADRAKGQFDSEGLMDLSKLEGAFSDPTMQFY
LCGPVGFMQFTAKQLVDLGVKQENIHYECFGPHKVL
```

Or, upload a file: Choose file No file chosen

AND

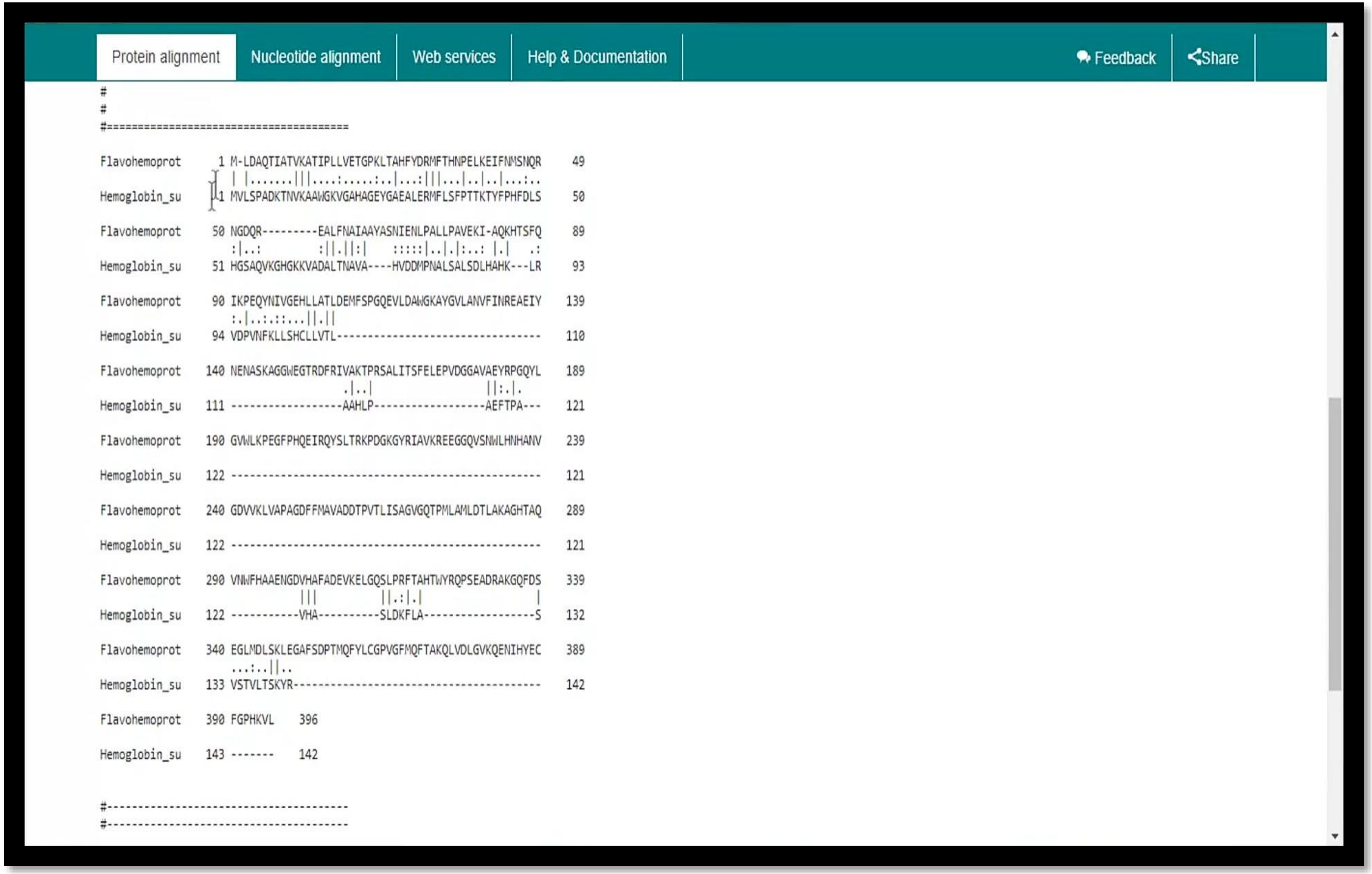
Enter or paste your second **protein** sequence in any supported format:

```
>Hemoglobin_subunit_alpha_Homo_sapiens
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR
```

Or, upload a file: Choose file No file chosen

STEP 2 - Set your pairwise alignment options

Entering the sequence to be compared in N-W algorithm



```
#
#
#=======================================

Flavohemoprot      1 M-LDAQTIATVKATIPLLVETGPKLTAHFYDRMFTHNPELKEIFNMSNQR     49
                     | |.......|||....:.....:..|...:|||...|..|..|...:..
Hemoglobin_su      1 MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS     50

Flavohemoprot     50 NGDQR---------EALFNAIAAYASNIENLPALLPAVEKI-AQKHTSFQ     89
                     :|..:         :||.||:|    :::::|..|.|:..: |.|   .:
Hemoglobin_su     51 HGSAQVKGHGKKVADALTNAVA----HVDDMPNALSALSDLHAHK---LR     93

Flavohemoprot     90 IKPEQYNIVGEHLLATLDEMFSPGQEVLDAWGKAYGVLANVFINREAEIY    139
                     :.|..:.::...||.||
Hemoglobin_su     94 VDPVNFKLLSHCLLVTL---------------------------------    110

Flavohemoprot    140 NENASKAGGWEGTRDFRIVAKTPRSALITSFELEPVDGGAVAEYRPGQYL    189
                                       .|..|                  ||:.|.
Hemoglobin_su    111 ------------------AAHLP------------------AEFTPA---    121

Flavohemoprot    190 GVWLKPEGFPHQEIRQYSLTRKPDGKGYRIAVKREEGGQVSNWLHNHANV    239

Hemoglobin_su    122 --------------------------------------------------    121

Flavohemoprot    240 GDVVKLVAPAGDFFMAVADDTPVTLISAGVGQTPMLAMLDTLAKAGHTAQ    289

Hemoglobin_su    122 --------------------------------------------------    121

Flavohemoprot    290 VNWFHAAENGDVHAFADEVKELGQSLPRFTAHTWYRQPSEADRAKGQFDS    339
                                |||          ||.:|.|                  |
Hemoglobin_su    122 -----------VHA----------SLDKFLA------------------S    132

Flavohemoprot    340 EGLMDLSKLEGAFSDPTMQFYLCGPVGFMQFTAKQLVDLGVKQENIHYEC    389
                     ...:..||..
Hemoglobin_su    133 VSTVLTSKYR----------------------------------------    142

Flavohemoprot    390 FGPHKVL    396

Hemoglobin_su    143 -------    142


#---------------------------------------
#---------------------------------------
```

Protein alignment result

Pairwise sequence alignment

- When comparing 2 sequences it is Pairwise sequence alignment (nucleic acids or protein)

- When comparing more than 2 sequence it is Mulitiple sequence alignment (nucleic acids or protein)

Contd…

- Pairwise sequence alignment is concerned with comparing 2 DNA or 2 Amino acids sequences

- For ex.

  BLASTn :- for nucleotide sequence

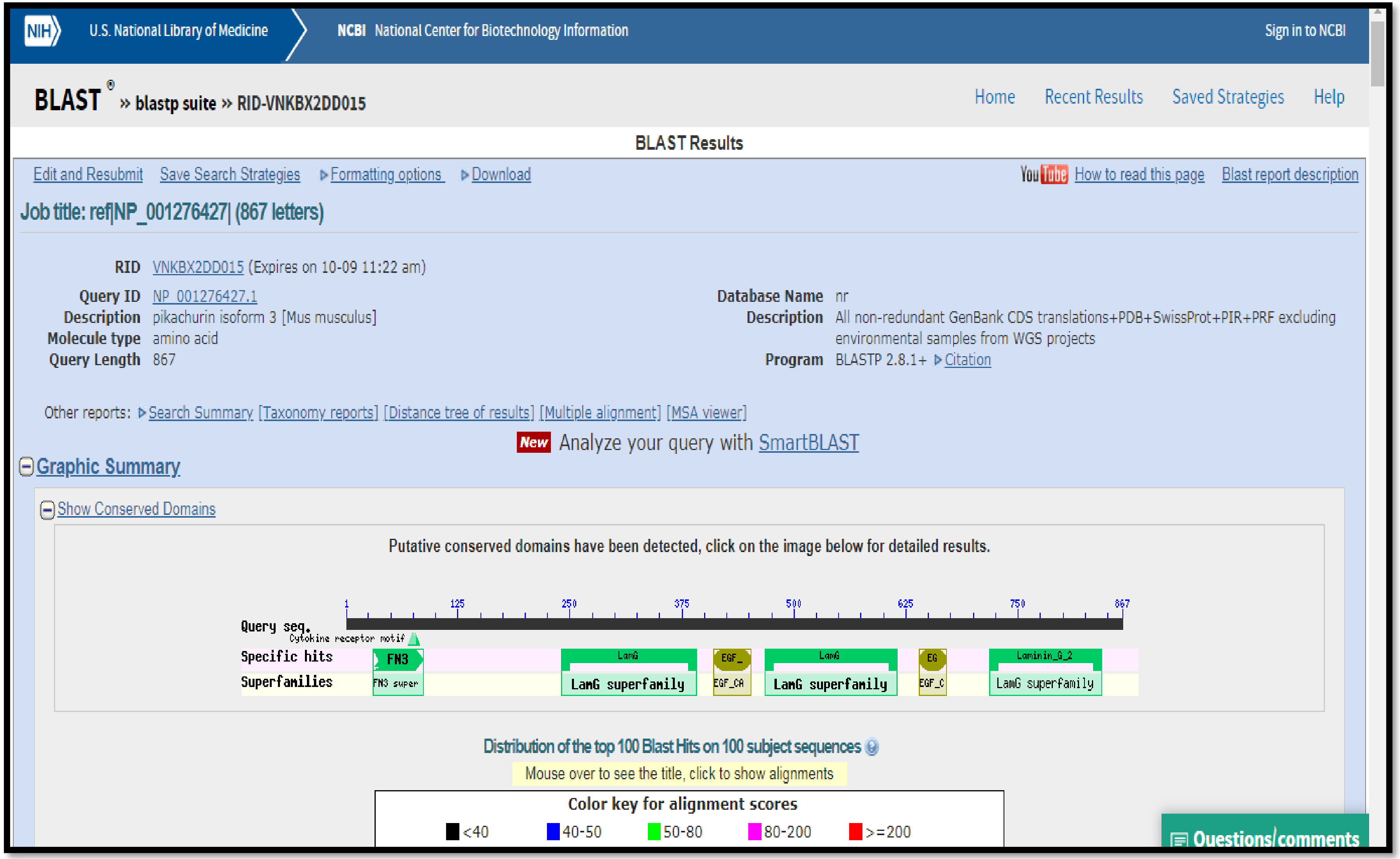
  BLASTp :- for protein sequence

BLAST-n and BLAST-P

- BLAST is a **B**asic **L**ocal **A**lignment **S**earch **T**ool.

- Used for comparing primary biological information viz,
  - Amino acid sequences of protein
  - Nucleotide of DNA or RNA sequences

- BLASTn and BLASTp is particularly used for comparing nucleotide sequence and protein sequence respectively



NIH U.S. National Library of Medicine

NCBI National Center for Biotechnology Information

Sign in to NCBI

BLAST® » blastp suite » RID-VNKBX2DD015

Home Recent Results Saved Strategies Help

BLAST Results

Edit and Resubmit Save Search Strategies ▷Formatting options ▷Download

YouTube How to read this page Blast report description

Job title: ref|NP_001276427| (867 letters)

RID VNKBX2DD015 (Expires on 10-09 11:22 am)

Query ID	NP_001276427.1	Database Name	nr
Description	pikachurin isoform 3 [Mus musculus]	Description	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule type	amino acid		
Query Length	867	Program	BLASTP 2.8.1+ ▷Citation

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] [MSA viewer]

New Analyze your query with SmartBLAST

⊖Graphic Summary

⊖Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of the top 100 Blast Hits on 100 subject sequences

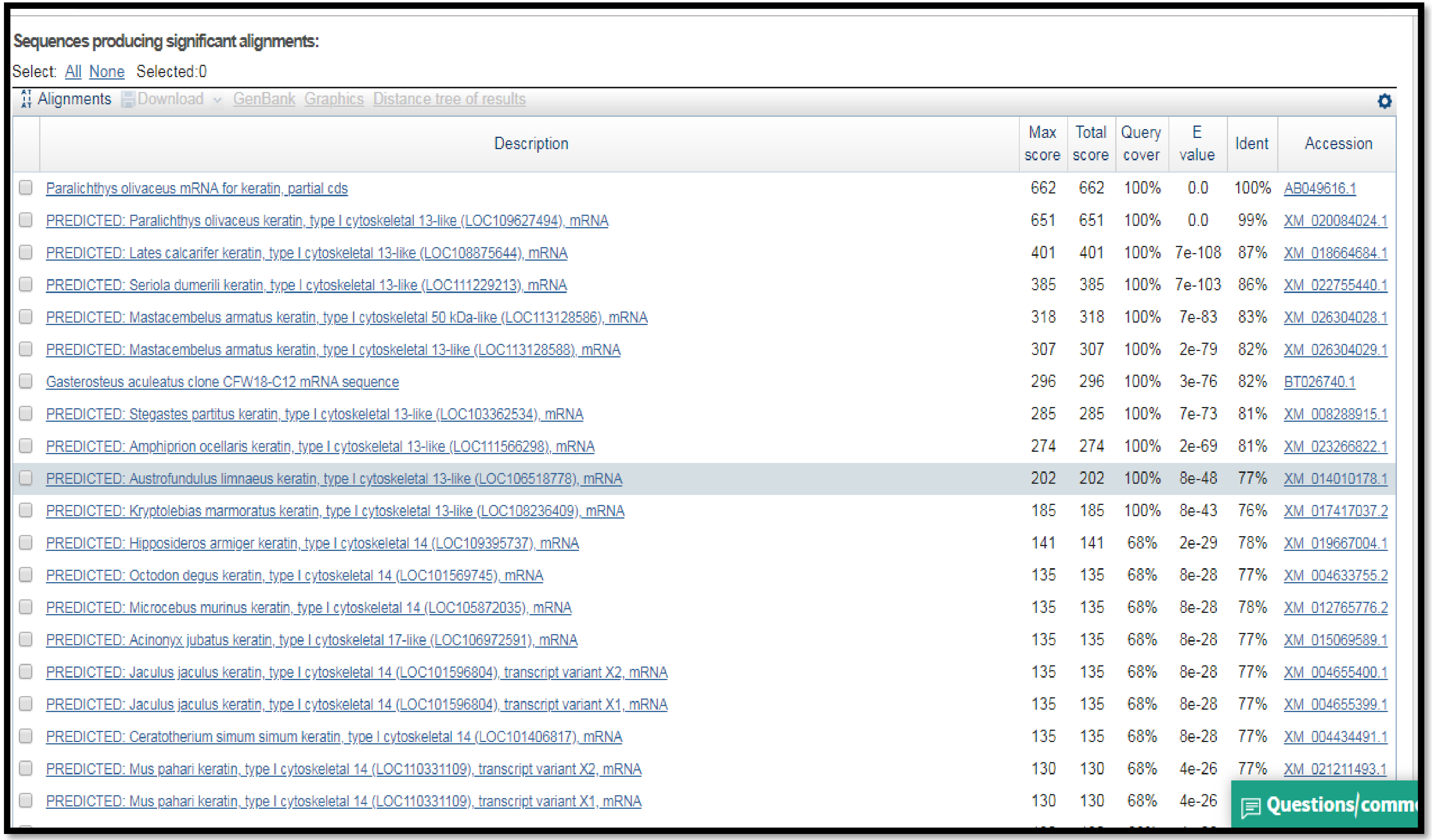
Mouse over to see the title, click to show alignments

Color key for alignment scores

■ <40 🟦 40-50 🟩 50-80 ■ 80-200 🟥 >=200

Questions/comments

Result of *Paralichthys olivaceus* in BLASTn

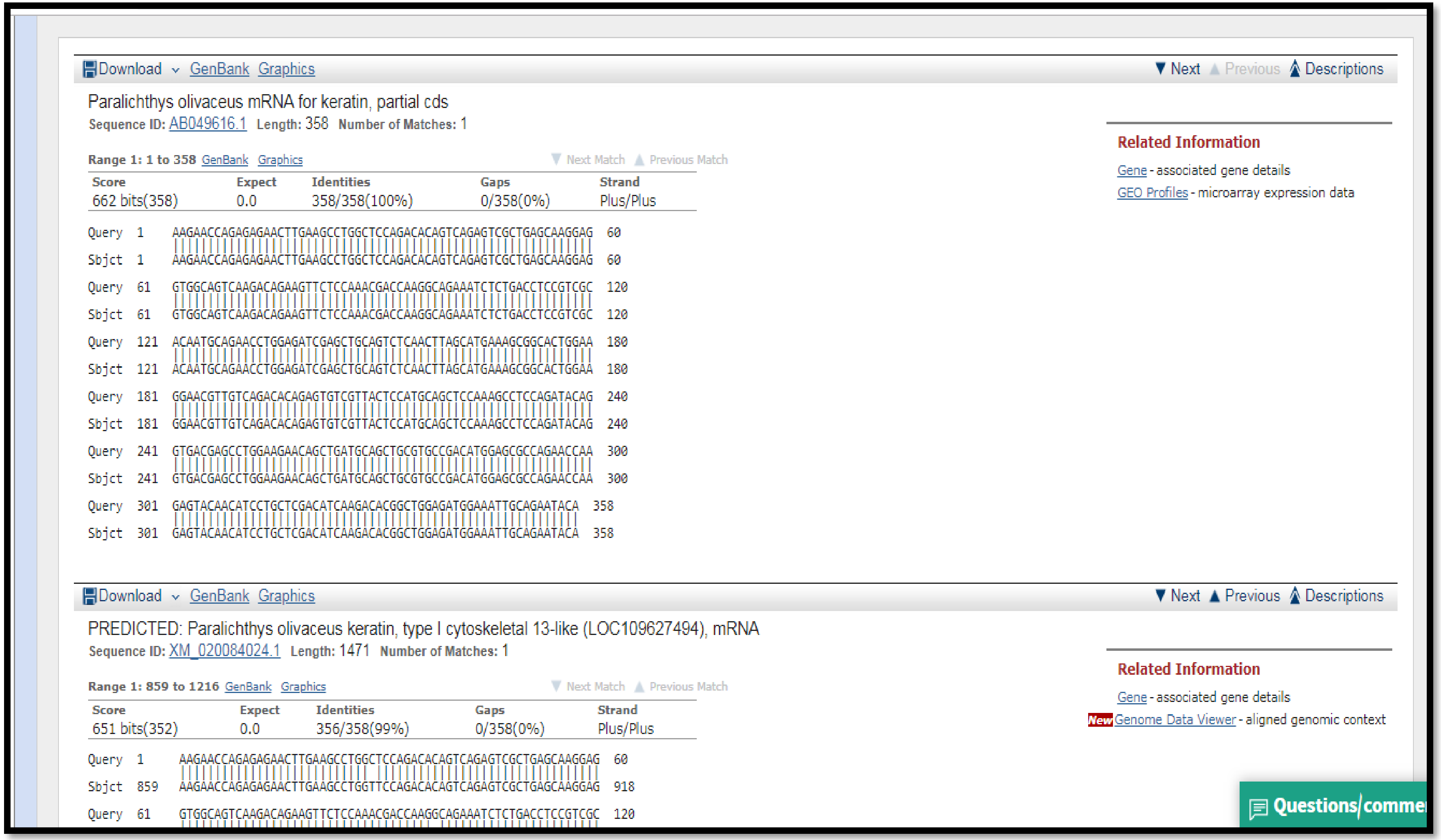


Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	Paralichthys olivaceus mRNA for keratin, partial cds	662	662	100%	0.0	100%	AB049616.1
□	PREDICTED: Paralichthys olivaceus keratin, type I cytoskeletal 13-like (LOC109627494), mRNA	651	651	100%	0.0	99%	XM_020084024.1
□	PREDICTED: Lates calcarifer keratin, type I cytoskeletal 13-like (LOC108875644), mRNA	401	401	100%	7e-108	87%	XM_018664684.1
□	PREDICTED: Seriola dumerili keratin, type I cytoskeletal 13-like (LOC111229213), mRNA	385	385	100%	7e-103	86%	XM_022755440.1
□	PREDICTED: Mastacembelus armatus keratin, type I cytoskeletal 50 kDa-like (LOC113128586), mRNA	318	318	100%	7e-83	83%	XM_026304028.1
□	PREDICTED: Mastacembelus armatus keratin, type I cytoskeletal 13-like (LOC113128588), mRNA	307	307	100%	2e-79	82%	XM_026304029.1
□	Gasterosteus aculeatus clone CFW18-C12 mRNA sequence	296	296	100%	3e-76	82%	BT026740.1
□	PREDICTED: Stegastes partitus keratin, type I cytoskeletal 13-like (LOC103362534), mRNA	285	285	100%	7e-73	81%	XM_008288915.1
□	PREDICTED: Amphiprion ocellaris keratin, type I cytoskeletal 13-like (LOC111566298), mRNA	274	274	100%	2e-69	81%	XM_023266822.1
□	PREDICTED: Austrofundulus limnaeus keratin, type I cytoskeletal 13-like (LOC106518778), mRNA	202	202	100%	8e-48	77%	XM_014010178.1
□	PREDICTED: Kryptolebias marmoratus keratin, type I cytoskeletal 13-like (LOC108236409), mRNA	185	185	100%	8e-43	76%	XM_017417037.2
□	PREDICTED: Hipposideros armiger keratin, type I cytoskeletal 14 (LOC109395737), mRNA	141	141	68%	2e-29	78%	XM_019667004.1
□	PREDICTED: Octodon degus keratin, type I cytoskeletal 14 (LOC101569745), mRNA	135	135	68%	8e-28	77%	XM_004633755.2
□	PREDICTED: Microcebus murinus keratin, type I cytoskeletal 14 (LOC105872035), mRNA	135	135	68%	8e-28	78%	XM_012765776.2
□	PREDICTED: Acinonyx jubatus keratin, type I cytoskeletal 17-like (LOC106972591), mRNA	135	135	68%	8e-28	77%	XM_015069589.1
□	PREDICTED: Jaculus jaculus keratin, type I cytoskeletal 14 (LOC101596804), transcript variant X2, mRNA	135	135	68%	8e-28	77%	XM_004655400.1
□	PREDICTED: Jaculus jaculus keratin, type I cytoskeletal 14 (LOC101596804), transcript variant X1, mRNA	135	135	68%	8e-28	77%	XM_004655399.1
□	PREDICTED: Ceratotherium simum simum keratin, type I cytoskeletal 14 (LOC101406817), mRNA	135	135	68%	8e-28	77%	XM_004434491.1
□	PREDICTED: Mus pahari keratin, type I cytoskeletal 14 (LOC110331109), transcript variant X2, mRNA	130	130	68%	4e-26	77%	XM_021211493.1
□	PREDICTED: Mus pahari keratin, type I cytoskeletal 14 (LOC110331109), transcript variant X1, mRNA	130	130	68%	4e-26		

Sequence producing significant alignment



Paralichthys olivaceus mRNA for keratin, partial cds

Sequence ID: AB049616.1 **Length:** 358 **Number of Matches:** 1

Range 1: 1 to 358 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
662 bits(358)	0.0	358/358(100%)	0/358(0%)	Plus/Plus

```
Query  1    AAGAACCAGAGAGAACTTGAAGCCTGGCTCCAGACACAGTCAGAGTCGCTGAGCAAGGAG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    AAGAACCAGAGAGAACTTGAAGCCTGGCTCCAGACACAGTCAGAGTCGCTGAGCAAGGAG  60

Query  61   GTGGCAGTCAAGACAGAAGTTCTCCAAACGACCAAGGCAGAAATCTCTGACCTCCGTCGC  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   GTGGCAGTCAAGACAGAAGTTCTCCAAACGACCAAGGCAGAAATCTCTGACCTCCGTCGC  120

Query  121  ACAATGCAGAACCTGGAGATCGAGCTGCAGTCTCAACTTAGCATGAAAGCGGCACTGGAA  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  ACAATGCAGAACCTGGAGATCGAGCTGCAGTCTCAACTTAGCATGAAAGCGGCACTGGAA  180

Query  181  GGAACGTTGTCAGACACAGAGTGTCGTTACTCCATGCAGCTCCAAAGCCTCCAGATACAG  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  GGAACGTTGTCAGACACAGAGTGTCGTTACTCCATGCAGCTCCAAAGCCTCCAGATACAG  240

Query  241  GTGACGAGCCTGGAAGAACAGCTGATGCAGCTGCGTGCCGACATGGAGCGCCAGAACCAA  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  GTGACGAGCCTGGAAGAACAGCTGATGCAGCTGCGTGCCGACATGGAGCGCCAGAACCAA  300

Query  301  GAGTACAACATCCTGCTCGACATCAAGACACGGCTGGAGATGGAAATTGCAGAATACA  358
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  GAGTACAACATCCTGCTCGACATCAAGACACGGCTGGAGATGGAAATTGCAGAATACA  358
```

PREDICTED: Paralichthys olivaceus keratin, type I cytoskeletal 13-like (LOC109627494), mRNA

Sequence ID: XM_020084024.1 **Length:** 1471 **Number of Matches:** 1

Range 1: 859 to 1216 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
651 bits(352)	0.0	356/358(99%)	0/358(0%)	Plus/Plus

```
Query  1    AAGAACCAGAGAGAACTTGAAGCCTGGCTCCAGACACAGTCAGAGTCGCTGAGCAAGGAG  60
            ||||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  859  AAGAACCAGAGAGAACTTGAAGCCTGGTTCCAGACACAGTCAGAGTCGCTGAGCAAGGAG  918

Query  61   GTGGCAGTCAAGACAGAAGTTCTCCAAACGACCAAGGCAGAAATCTCTGACCTCCGTCGC  120
```

Related Information

Gene - associated gene details

GEO Profiles - microarray expression data

Related Information

Gene - associated gene details

New Genome Data Viewer - aligned genomic context

Sequence alignment of *Paralichthys olivaceus*

Applications of pairwise sequence alignment

✓ Searching large sequences for matches

✓ Characterize newly sequenced genes or gene products

✓ Molecular distance of evolution between species

Lunch Break

Day 2: Sequence Alignment and Phylogenetic

Afternoon Session

Outline Day 2

Afternoon (2-5 pm): Phylogentics

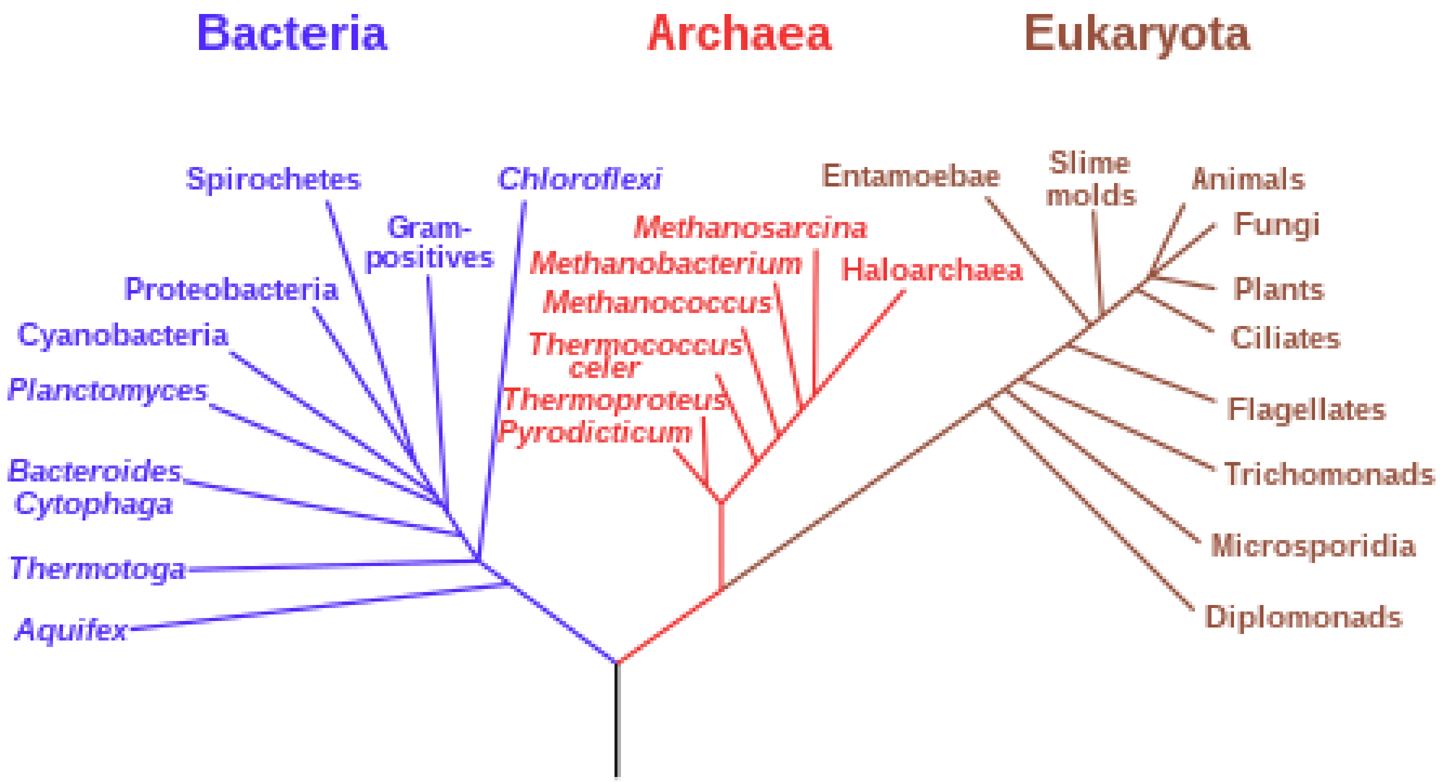
Phylogentics:

- Introduction to Phylogentics tree concept, and significance in evolution biology.

Practical Session:

- Construct phylogentics trees using Clustal Omega (web-based) starting from multiple sequence alignments.
- Code with python

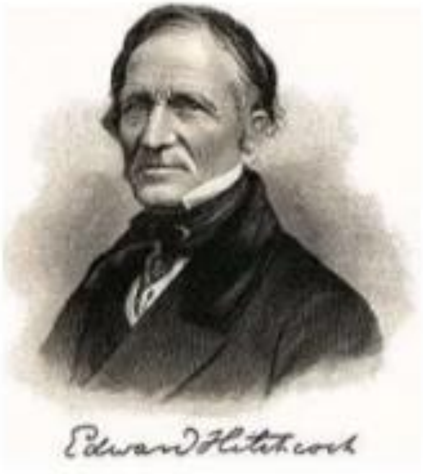
Phylogenetic tree



- A **phylogenetic tree** is a diagram that represents evolutionary relationships among organisms based on the similarities and differences in their genetic and evolutionary characteristics
- The pattern of branching in a phylogenetic tree reflects how species or other groups evolved from a series of common ancestors.
- The phylogenetic tree is also called the “Tree of Life” or “Dendrogram”

History

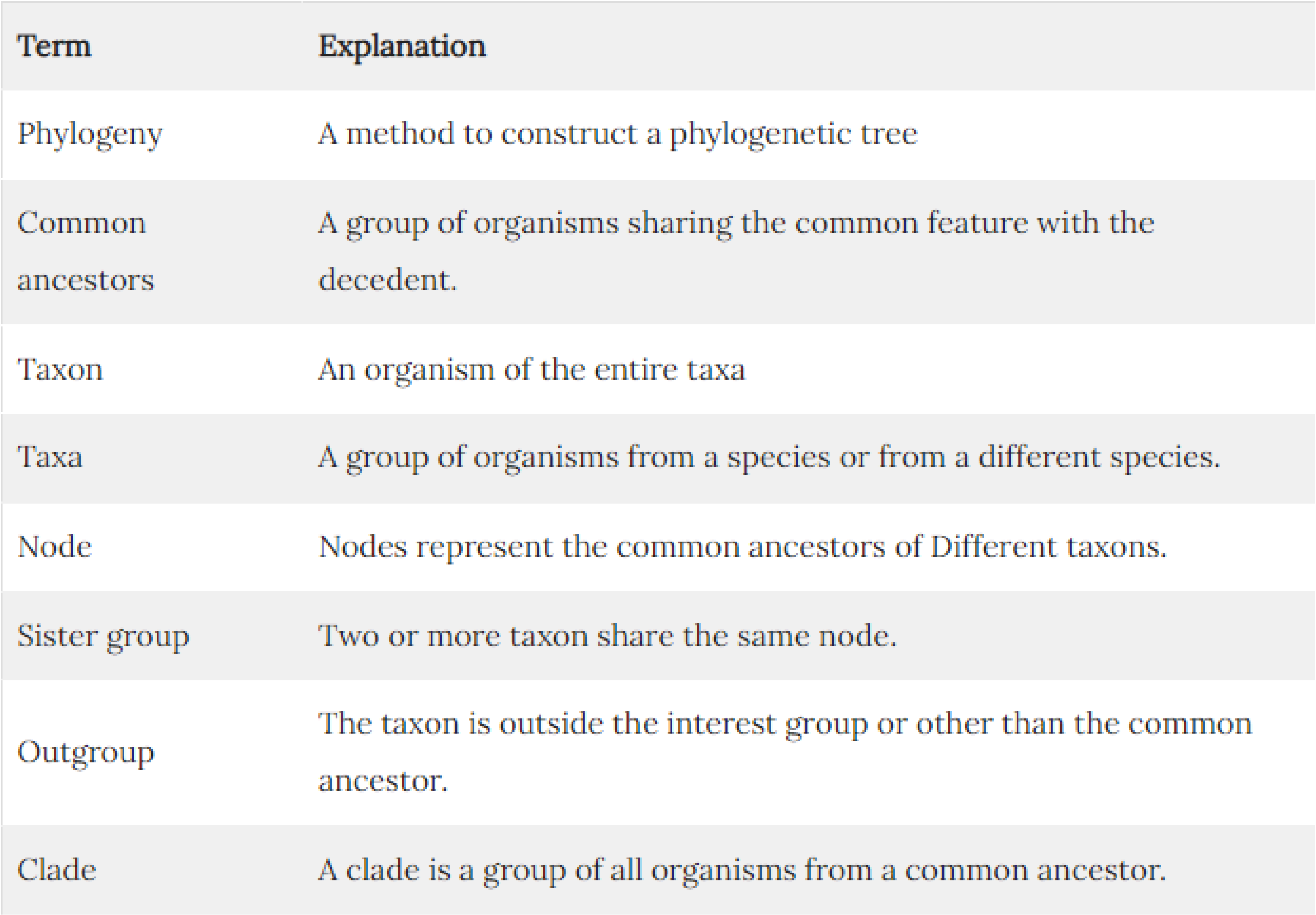
- Early representations of "branching" phylogenetic trees include a "paleontological chart" showing the geological relationships among plants and animals in the book *Elementary Geology*, by Edward Hitchcock in 1840.



- Charles Darwin in 1859 also produced one of the first illustrations and crucially popularized the notion of an evolutionary "tree" in his seminal book *The Origin of Species*.

Importance of Phylogenetic Tree

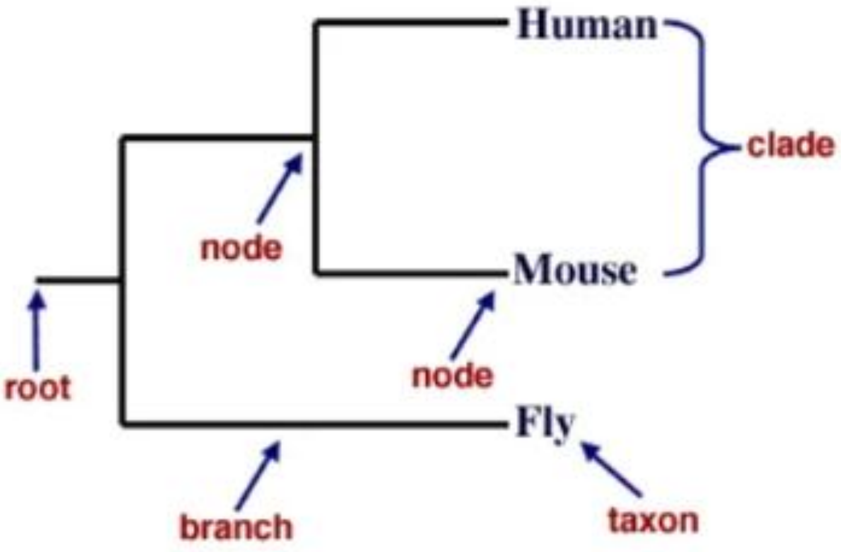
- It is the fundamental tool to derive their most-useful evidence from the fields of anatomy, embryology, palaeontology and molecular genetics. Other significances of the phylogenetic tree are:
- Used in the search for a new species.
- Used to study evolutionary histories.
- To study how the species were spread geographically.
- To study the common ancestors of extant and extinct species.
- It is used to identify the most recent common ancestors and to recognize how closely related species are.
- To relate the milestones of the evolution of major life forms to the tree of life.
- To represent evolutionary relationships between organisms that are believed to have some common ancestry.
- With the help of the phylogenetic tree, the infectious microbes can be traced along with their evolutionary histories.

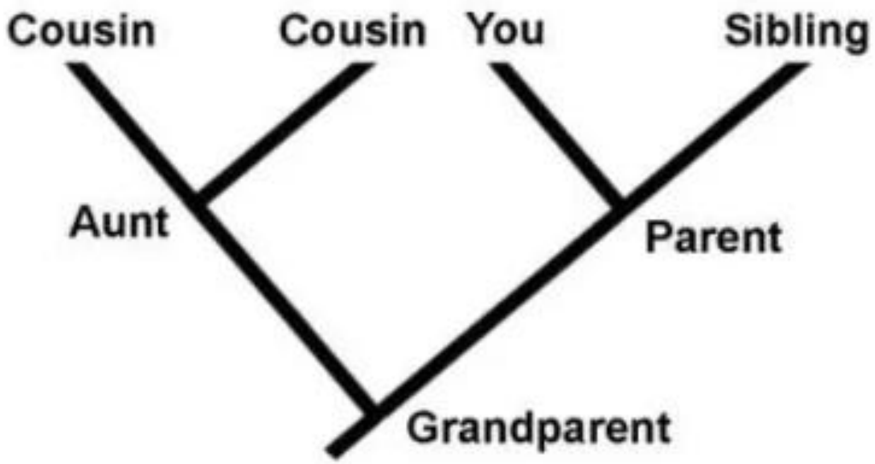


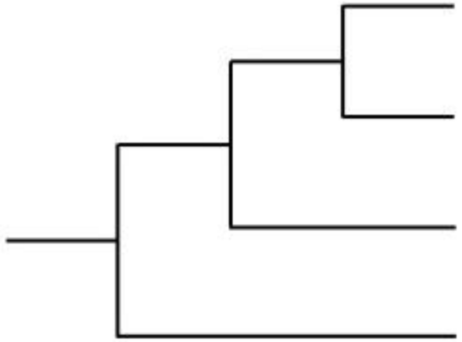
Term	Explanation
Phylogeny	A method to construct a phylogenetic tree
Common ancestors	A group of organisms sharing the common feature with the decedent.
Taxon	An organism of the entire taxa
Taxa	A group of organisms from a species or from a different species.
Node	Nodes represent the common ancestors of Different taxons.
Sister group	Two or more taxon share the same node.
Outgroup	The taxon is outside the interest group or other than the common ancestor.
Clade	A clade is a group of all organisms from a common ancestor.

What does this tree looks like?

- ➢What do the lines represent?

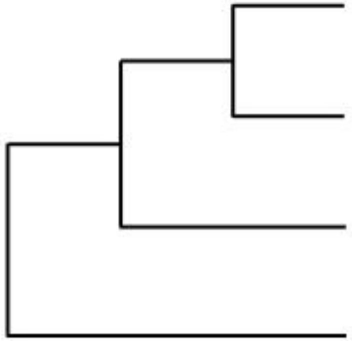






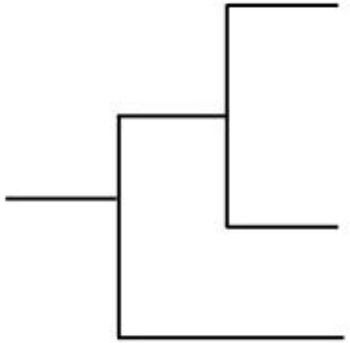
A

Rooted tree



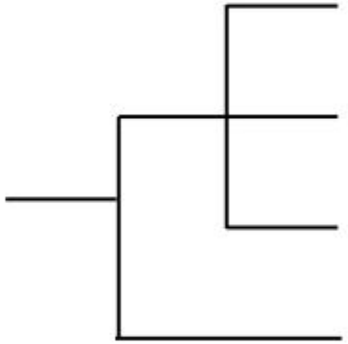
B

Unrooted tree



C

Bifurcating tree

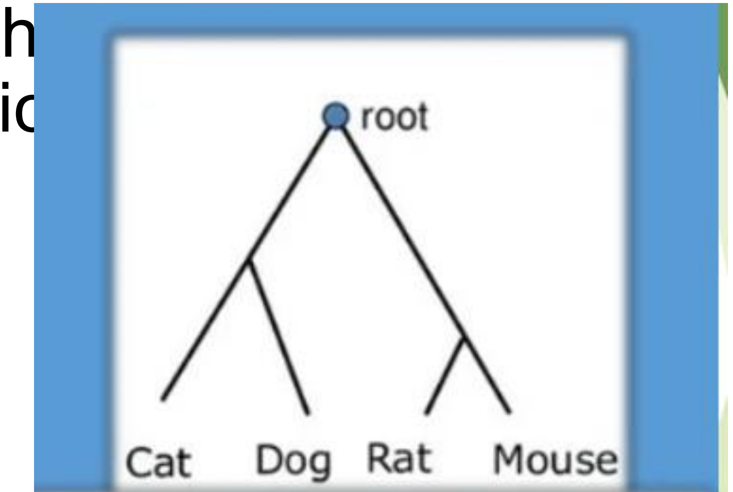


D

Multifurcating tree

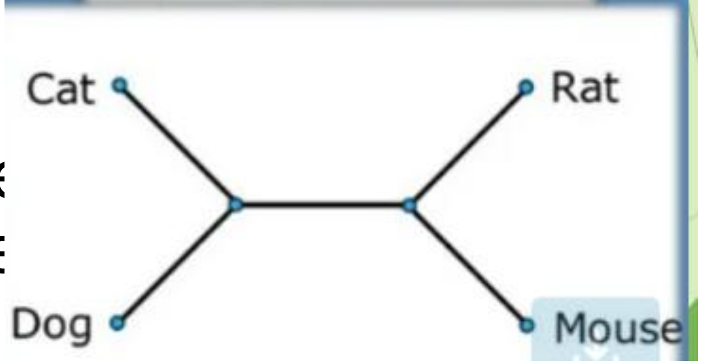
Rooted tree:

- The rooted tree is described as a phylogenetic tree sharing the common ancestor on the node. Therefore th classification ends at one point usually on the node whic is the common ancestor of all the branches of the tree.



Unrooted tree:

- Contrary to the rooted tree, the non-rooted tree doesn’t have a common ancestor. The unrooted phylogenetic tre is always prepared from the rooted tree by excluding the common ancestor or the node of the tree.

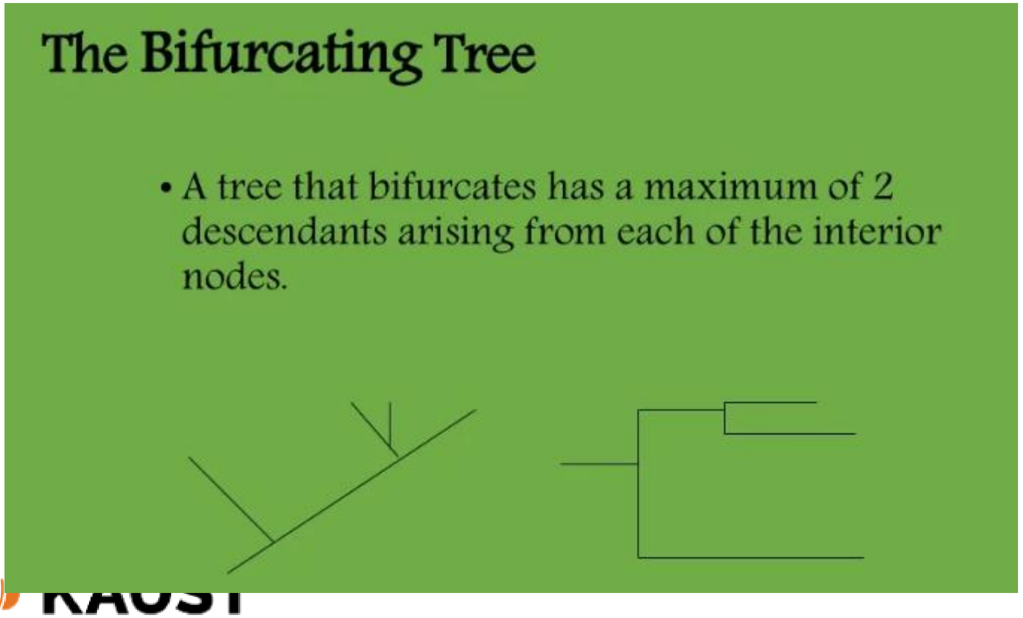


Bifurcating tree

- The phylogenetic tree only has two branches or we can say leaves are known as bifurcating trees. It is also classified in rooted bifurcation trees and unrooted bifurcating trees.

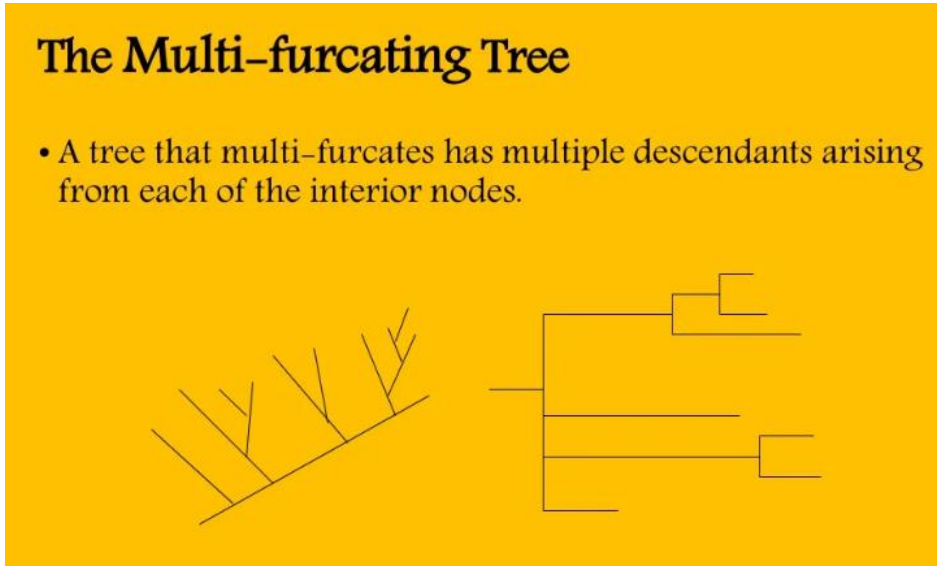
Multifurcating tree:

- The multifurcating tree is described as having multiple branches on a single node. Again, it is classified into a rooted multifurcating tree and an unrooted multifurcating tree.



The Bifurcating Tree

- A tree that bifurcates has a maximum of 2 descendants arising from each of the interior nodes.

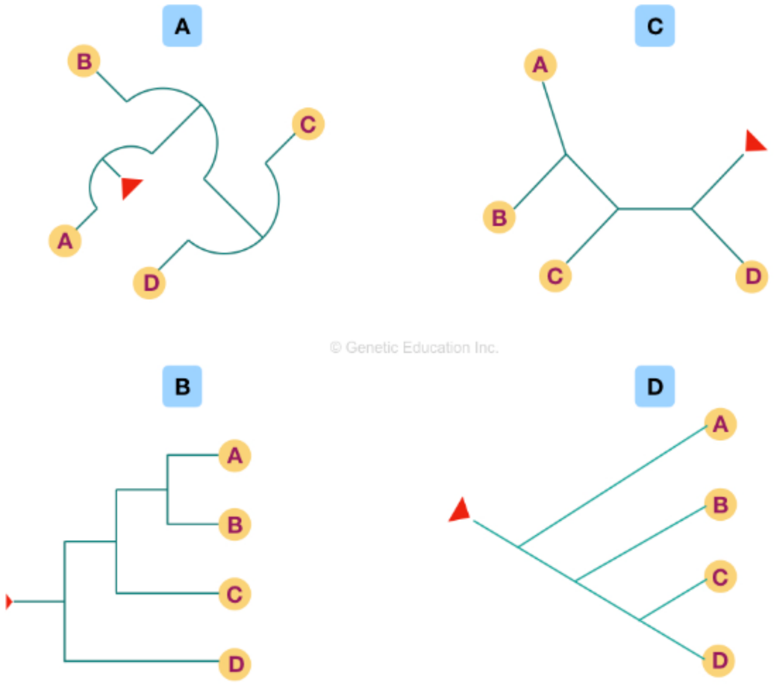


The Multi-furcating Tree

- A tree that multi-furcates has multiple descendants arising from each of the interior nodes.

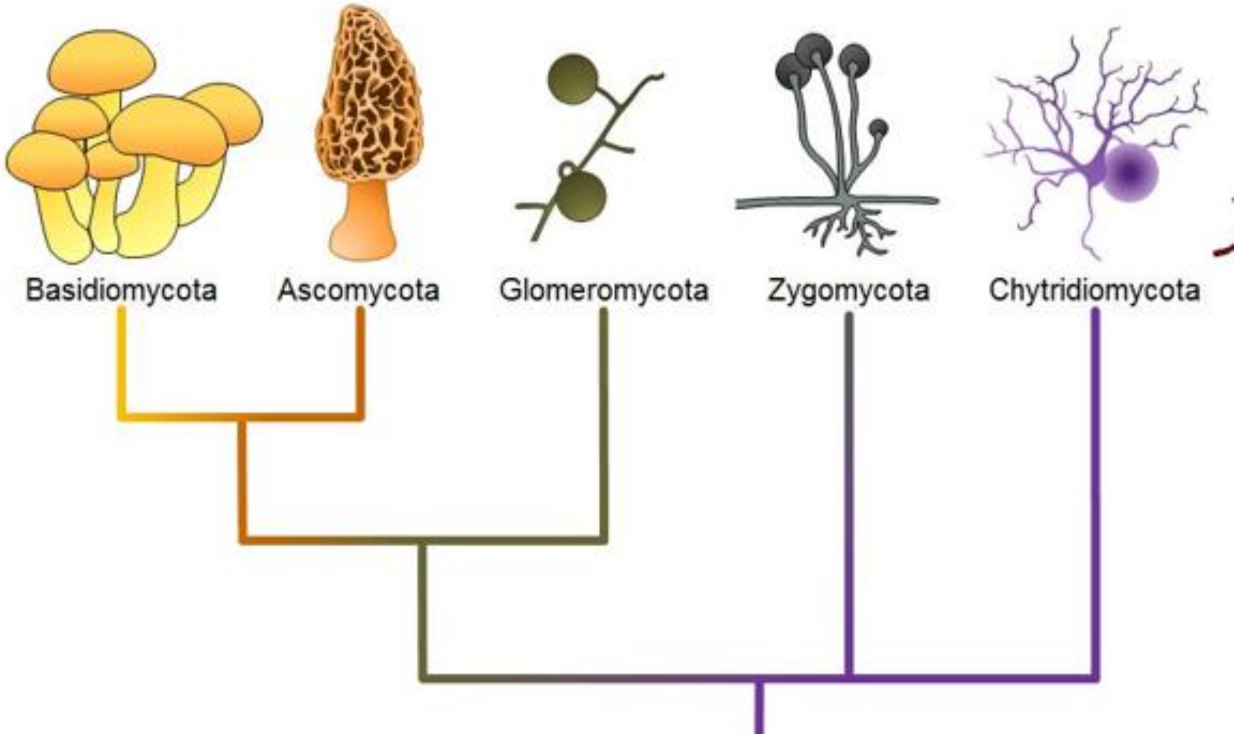
In phylogeny, the node is also known as a “clade” as well. Though there are so many different variations of the phylogenetic tree, every method of making a tree depicts the same type of information. Take a look at various trees shown in the figure

Keep in mind that whatever the shape, topology or structure of the tree is, it must have a common node if rooted and branched. To read a tree, start with the tip of the branches and see where the branch ends (the node), based on that information you can depict or conclude which organism is nearer or closer and which are distantly related.



The figure represents different forms of a single type of phylogenetic tree.

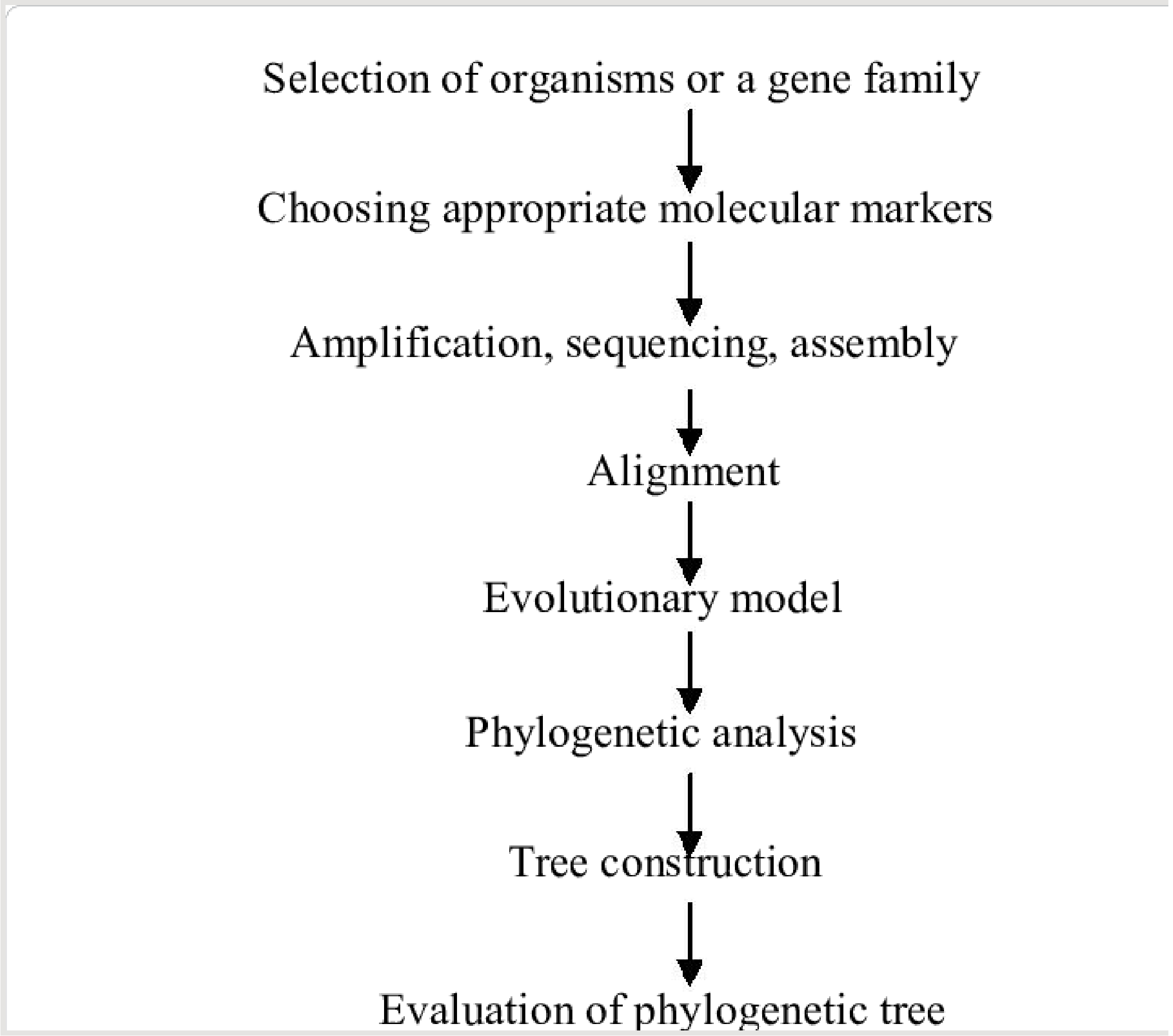
Phylogeny in fungus



Applications of a phylogenetic tree:

- The phylogenetic tree is constructed to make an evolutionary link between various organisms. By doing so, we can get an idea about how and from whom different organisms are evolved.
- Also, it helps to classify organisms and species in different taxa and groups based on their DNA sequence and phenotypic similarities and differences.
- In addition to this, it is useful to study the force of evolution and characteristics of different organisms.
- it is applicable to study the events occurring during the course of evolution and to classify species based on the divergence of structure and function.

Steps in Phylogenetic Analysis



Selection of organisms or a gene family

Choosing appropriate molecular markers

Amplification, sequencing, assembly

Alignment

Evolutionary model

Phylogenetic analysis

Tree construction

Evaluation of phylogenetic tree

Software

- Some programs for phylogenetic analysis
- A multiple alignment program:
- Clustal, T-Coffee, MAFFT, Muscle…
- A phylogenetic program:
- Phylip, PAUP*, MacClade, BioEdit…
- Visualizing the tree:
- TreeView, Njplot
- **https://evolution.genetics.washington.edu/phylip/software.html**

Selecting sequence

- The rate of mutation is assumed to be the same in both coding and non-coding region
- However there is a difference in substitution rate
- Non-coding DNA region have more substituion than coding regions.
- Protien are much more conserved since they need to conseve their function
- It is better to use sequence that mutate slowly (protein) than DNA. If the gene are very small or they mutate slowly, then it cam use for building tree.

Building Phylogenetic Trees

- The most popular and frequently used methods of tree building can be classified into two major categories
- Phenetic methods based on **distances**
- Cladistic methods based on **characters**
- **Distance matrix methods**
  - UPGMA (Unweighted pair group methods with arithmetic mean)
  - Fitech-Margoliash
  - Neighbour joining (NJ)
- **Character based methods**
  - Maximum parsimony (MP)
  - Maximum likelihood (ML)

Distance based methods

- Tree are calculated by similiarities of sequences and are based on distance
- Some sequences more similar than others
- Closely related sequences should be close in the tree
- Only use the distances between sequences
- All methods start with a *distance matrix*

UPGMA Vs Fitch Margoliash Method

UPGMA Methods

Unweighted Pair Group Method with Arithmetic Mean

Unweighted: The distances are used as they are

Pair: Find the two closest elements

Group: Put them together in a new group

Arithmetic Mean: Gives distances from the new group

Fitch-Margoliash Methods

More complicated than UPGMA

Does not assume a molecular clock

Produces an unrooted scaled tree

Neighbour joining

- This methods tries to correct the UPGMA method for its assumption that the rate of evolutions is the same in all taxa.
- But it assumes an additive tree
  - Distance between two leaves is the sum of the edges
- Find the *closest pair* that is *most apart* from the rest of the tree
- Connect pair and update distances
  - A little advanced: Take the overall distance to the rest of the tree into account
  - Corrects for varying mutation
- Fast and can give good results

Character methods

- Tree are calculated by considering the various possible pathways of evolution.
- This methods uses each alignment positions as evolutionary information to build a tree.
- All information at hand
- More advanced, slower, but also more accurate
- Maximum Parsimony (MP)
  - Occam’s razor: Simplest explanation
- Maximum Likelihood (ML)
  - Advanced statistical method
  - Most probable tree given the data and the model

Maximum parsimony (MP)

- For each positon in the alignemnt all possible trees are evaluated and are given a score based on the number of evooution changes.
- More time consuming
- Used for closely related sequences
- The most parsimonious tree is the one with the fewest evolutionary changes
- MP methods are available for DNA in Programs **paup, molphy, phylo_win**

Maximum Likelihood

- This methods also uses each position in an alignment, evaluate all possible trees and calcultes the likelihood for each tree.
- The tree with the maximum likelihood is the most probable tree.
- Slowest methods but gives the best result
- Used for any set of sequence.
- Maximum likelikhood methods can found in **phylip, paup or puzzle**