OVERVIEW

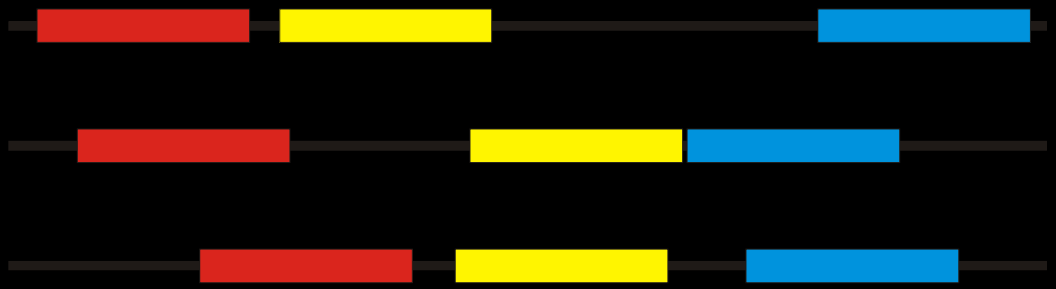
- Sequence alignment
- Types of sequence alignments
- Multiple sequence alignment
- Purpose of MSA
- Types of MSA
- Progressive alignment
- Tools

SEQUENCE ALIGNMENT

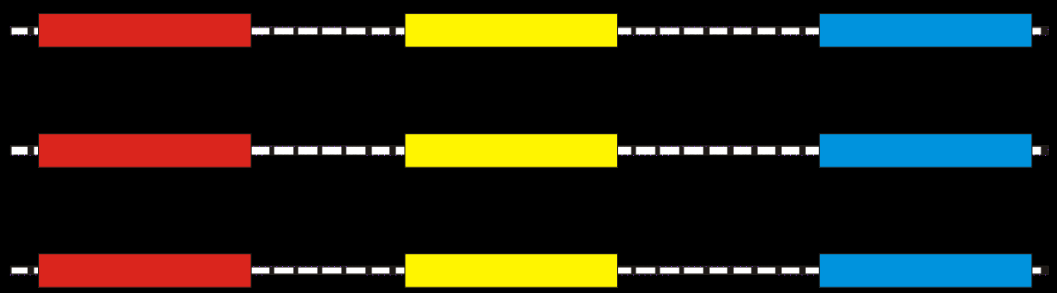
- In bioinformatics, a **sequence alignment** is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

SEQUENCE ALIGNMENT

Sequences often contain highly conserved regions



These regions can be used for initial alignments



TYPES OF SEQUENCE ALIGNMENTS

Pair- wise alignment

- Dot matrix method
- Dynamic programming
- Word methods

Multiple sequence alignment

- Dynamic programming
- Progressive methods
- Iterative methods

MULTIPLE SEQUENCE ALIGNMENT

- A multiple sequence alignment is an alignment of n > 2 sequences obtained by inserting gaps ("-") into sequences such that the resulting sequences have all length L and can be arranged in a matrix of N rows and L columns where each column represents a homologous position.

- The principle is that multiple alignments are achieved by successive application of pairwise methods.

PURPOSE OF MSA

- In order to characterize protein families, identify shared regions of homology in a multiple sequence alignment
- Determination of the consensus sequence of several aligned sequences.
- Consensus sequences can help to develop a sequence "finger print" which allows the identification of members of distantly related protein family (motifs)
- MSA can help us to reveal biological facts about proteins, like analysis of the secondary/tertiary structure

Main applications of multiple sequence alignments

Application	*Procedure*
Extrapolation	A good multiple alignment can help convincing you that an uncharacterized sequence is really a member of a protein family.
Phylogenetic analysis	If you carefully chose the sequences to include in your multiple alignment, you can reconstruct the history of these proteins.
Pattern Identification	By discovering very conserved positions you can identify a region that is characteristic of a function (in proteins or in nucleic acid sequences).
Domain identification	It is possible to turn a multiple sequence alignment into a profile that describes a protein family or a protein domain. You can use this profile to scan databases for new members of the family.
DNA regulatory elements	You can turn a DNA multiple alignment of a binding site into a weight matrix and scan other DNA sequences for potential similar binding sites.
Structure prediction	A good multiple alignment can give you an almost perfect prediction of your protein secondary structure for proteins or RNA. Sometimes it can also help building a 3-D model.
PCR analysis	A good multiple alignment can help you identifying the less degenerated portions of a protein family
nsSNP	Identify the nsSNP that are the most likely to alter the function

TYPES OF MSA

- Dynamic programming approach

  Computes an optimal alignment for a given score function. Because of its high running time , it is not typically used in practice.

- Progressive method

  This approach repeatedly aligns two sequences, two alignments, or a sequence with an alignment.

- Iterative method

  Works similarly to progressive methods but repeatedly realigns the initial sequences as well as adding new sequences to the growing MSA.

PROGRESSIVE ALIGNMENT

- The most widely used approach
- Builds up a final MSA by combining pairwise alignments beginning with the most similar pair and progressing to the most distantly related
- Progressive alignment methods require two stages:
  - A first stage in which the relationships between the sequences are represented as a tree, called a *guide tree*
  - Second step in which the MSA is built by adding the sequences sequentially to the growing MSA according to the guide tree

MSA USING CLUSTAL W

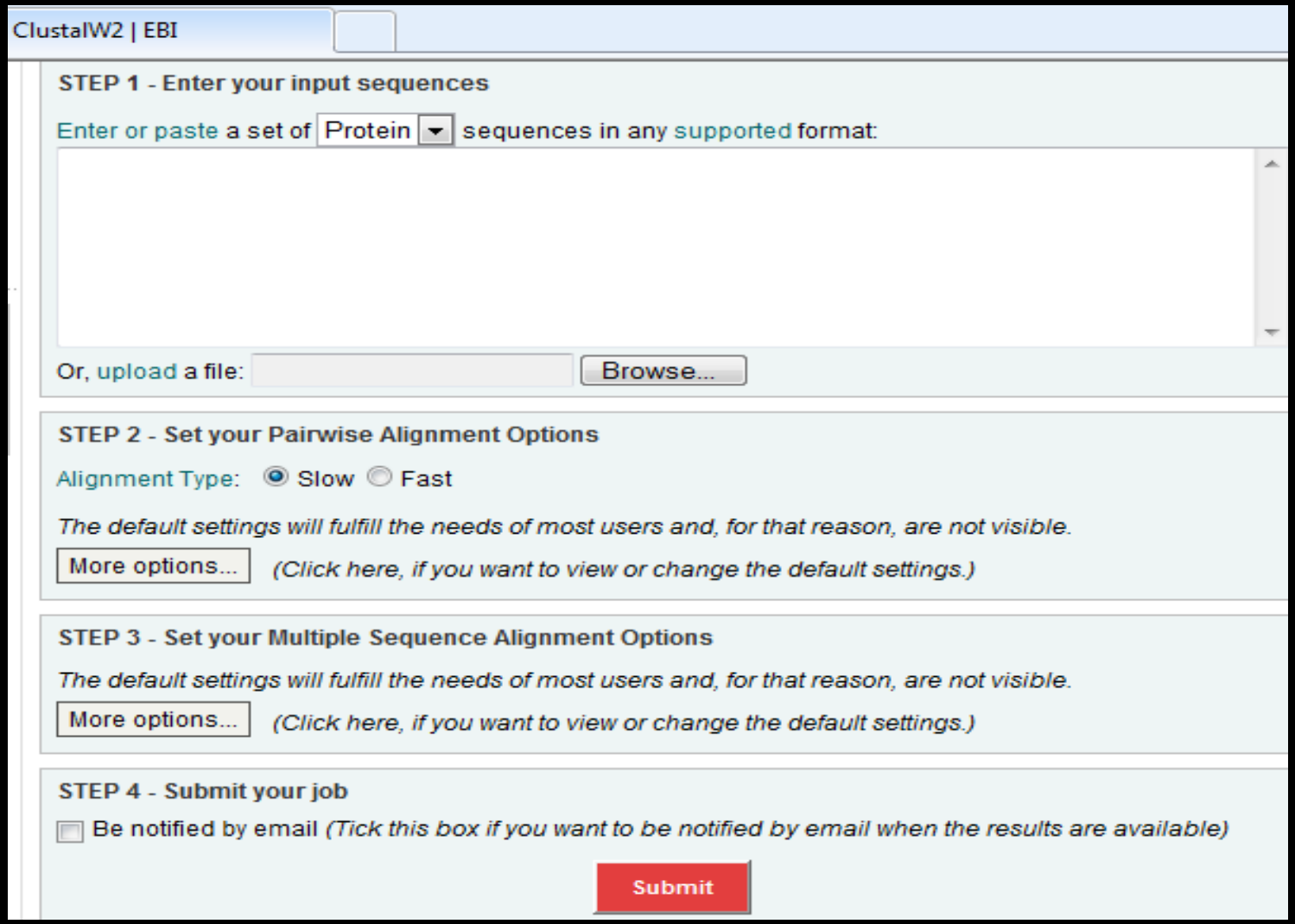
- Works by progressive alignment.
- Clustal W was introduced by Julie D. Thompson and Toby Gibson of EMBL, EBI.
- Most closely related sequences are aligned first, and then additional sequences and groups of sequences are added, guided by the initial alignments .
- Uses alignment scores to produce a phylogenetic tree.

MSA USING CLUSTAL W

- Aligns the sequences sequentially, guided by the phylogenetic relationships indicated by the tree.

- Gap penalties can be adjusted based on specific amino acid residues, regions of hydrophobicity, proximity to other gaps, or secondary structure.

CLUSTAL W


ClustalW2 | EBI
STEP 1 - Enter your input sequences
Enter or paste a set of Protein sequences in any supported format:
Or, upload a file:
Browse...
STEP 2 - Set your Pairwise Alignment Options
Alignment Type: Slow Fast
The default settings will fulfill the needs of most users and, for that reason, are not visible.
More options...
(Click here, if you want to view or change the default settings.)
STEP 3 - Set your Multiple Sequence Alignment Options
The default settings will fulfill the needs of most users and, for that reason, are not visible.
More options...
(Click here, if you want to view or change the default settings.)
STEP 4 - Submit your job
Be notified by email (Tick this box if you want to be notified by email when the results are available)
Submit

Overview of ClustalW Procedure

Hbb_Human	1	-			
Hbb_Horse	2	.17	-		
Hba_Human	3	.59	.60	-	
Hba_Horse	4	.59	.59	.13	-
Myg_Whale	5	.77	.77	.75	.75	-

alpha-helices

```
1  PEEKSAVTALWGKVN--VDEVGG
2  GEEKAAVLALWDKVN--EEEVGG
3  PADKTNVKAAWGKVGAHAGEYGA
4  AADKTNVKAAWSKVGGHAGEYGA
5  EHEWQLVLHVWAKVEADVAGHGQ
```

1 2 3 4

CLUSTAL W

↓

Quick pairwise alignment:
calculate distance matrix

↓

Neighbor-joining tree
(guide tree)

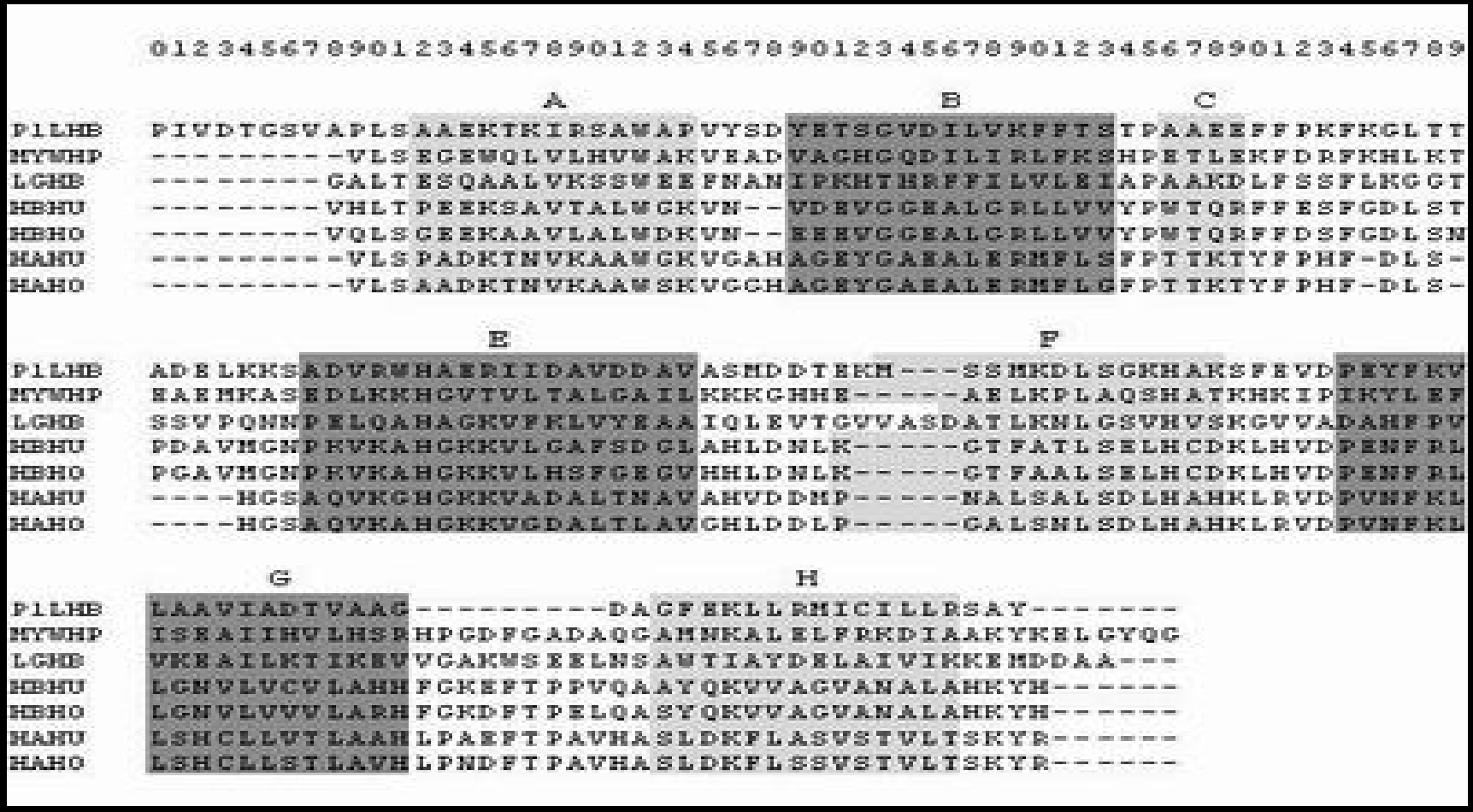
↓

Progressive alignment
following guide tree

WORKING OF CLUSTAL W

- **First perform all possible pairwise alignments between each pair of sequences.**
- **Calculate the 'distance' between each pair of sequences based on these isolated pairwise alignments.**
- **Generate a distance matrix.**
- **Generate a Neighbor-Joining 'guide tree' from these pairwise distances.**
- **This guide tree gives the order in which the progressive alignment will be carried out.**

IMAGE OF CLUSTAL W OUTPUT



MSA USING PILEUP

- It is the MSA program that is a part of the Genetics Computer Group Package of Sequence analysis program.
- The sequences are aligned pair wise using Needleman-Wunsch algorithm.
- The scores obtained are used to produce a tree by the UPGMA method.
- The resulting tree is then used to guide the alignment of the most closely related sequences.

CHOOSING SEQUENCES FOR PILEUP

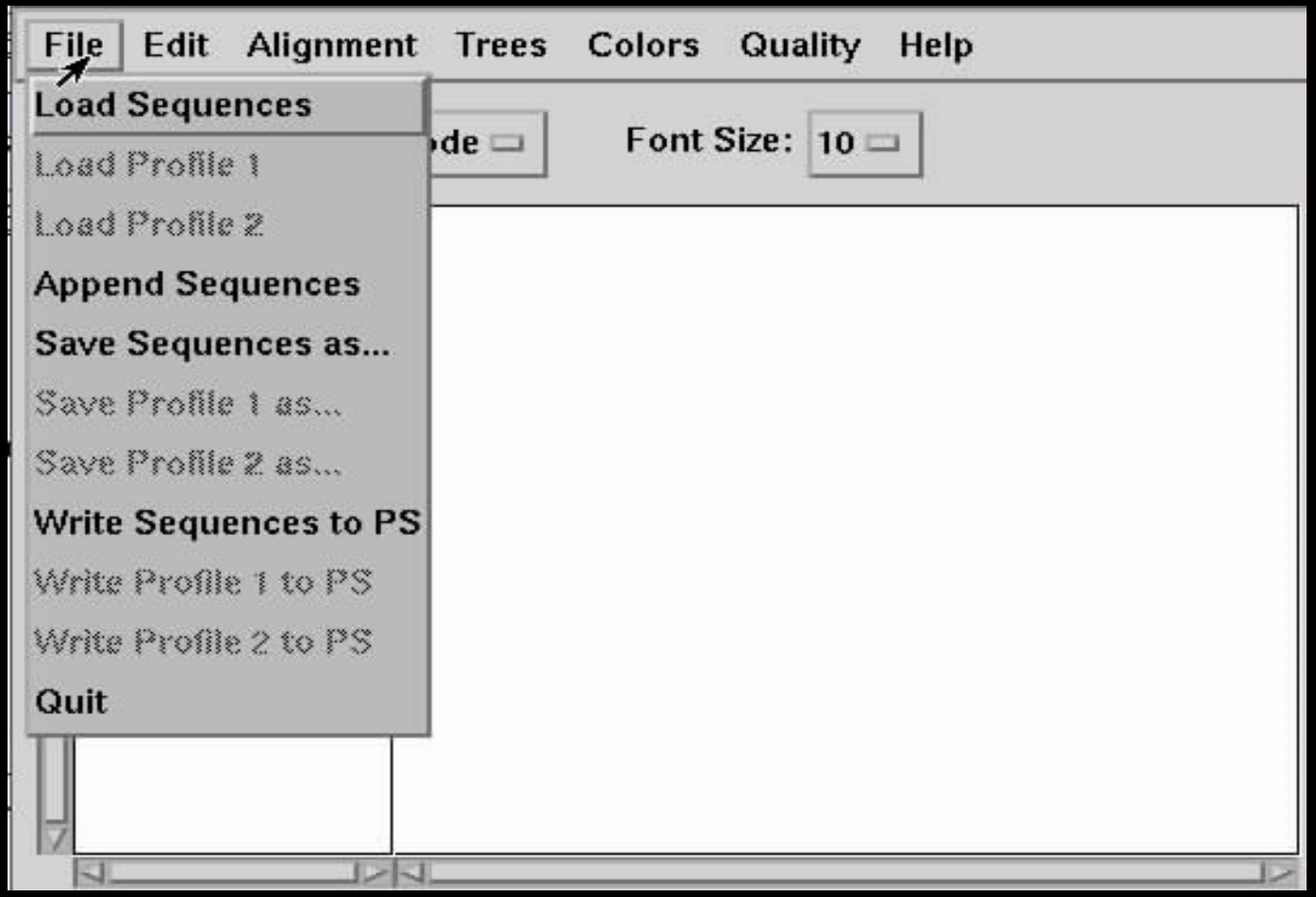
- As far as possible only the sequences of similar length are aligned.
- Pileup can align sequences of up to 5000 residues, with 2000 gaps (total 7000 characters).
- It is a good program only for similar (close) sequences.
- It does global multiple alignment, and therefore is good for a group of similar sequences

MSA USING CLUSTALX

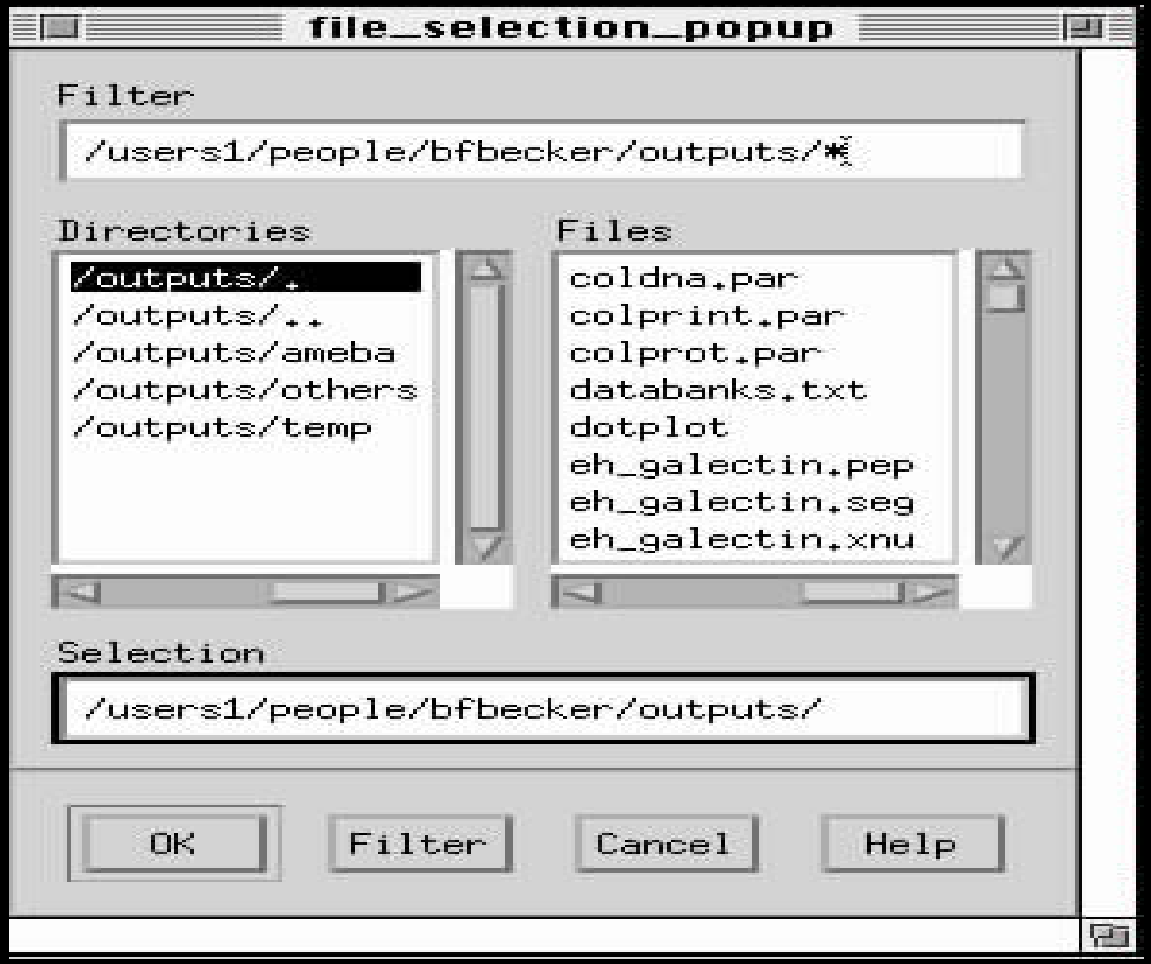
- ClustalX provides a new window-based user interface to the ClustalW program.

- It uses the Vibrant multi-platform user interface development library, developed by the National Center for Biotechnology as part of their NCBI software development toolkit

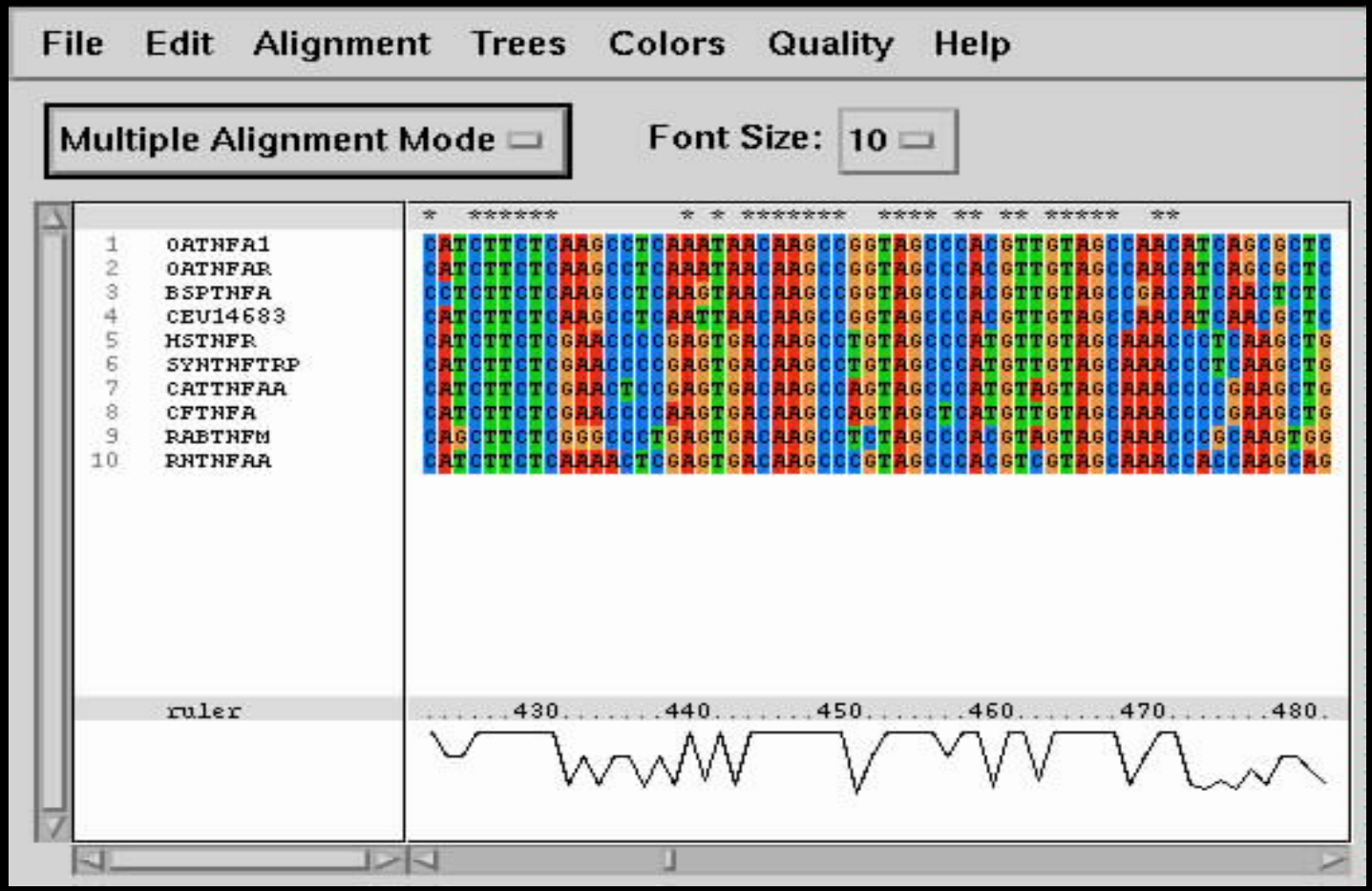
CLUSTALX


File
Edit
Alignment
Trees
Colors
Quality
Help
Load Sequences
Load Profile 1
Load Profile 2
Append Sequences
Save Sequences as...
Save Profile 1 as...
Save Profile 2 as...
Write Sequences to PS
Write Profile 1 to PS
Write Profile 2 to PS
Quit
Font Size:
10

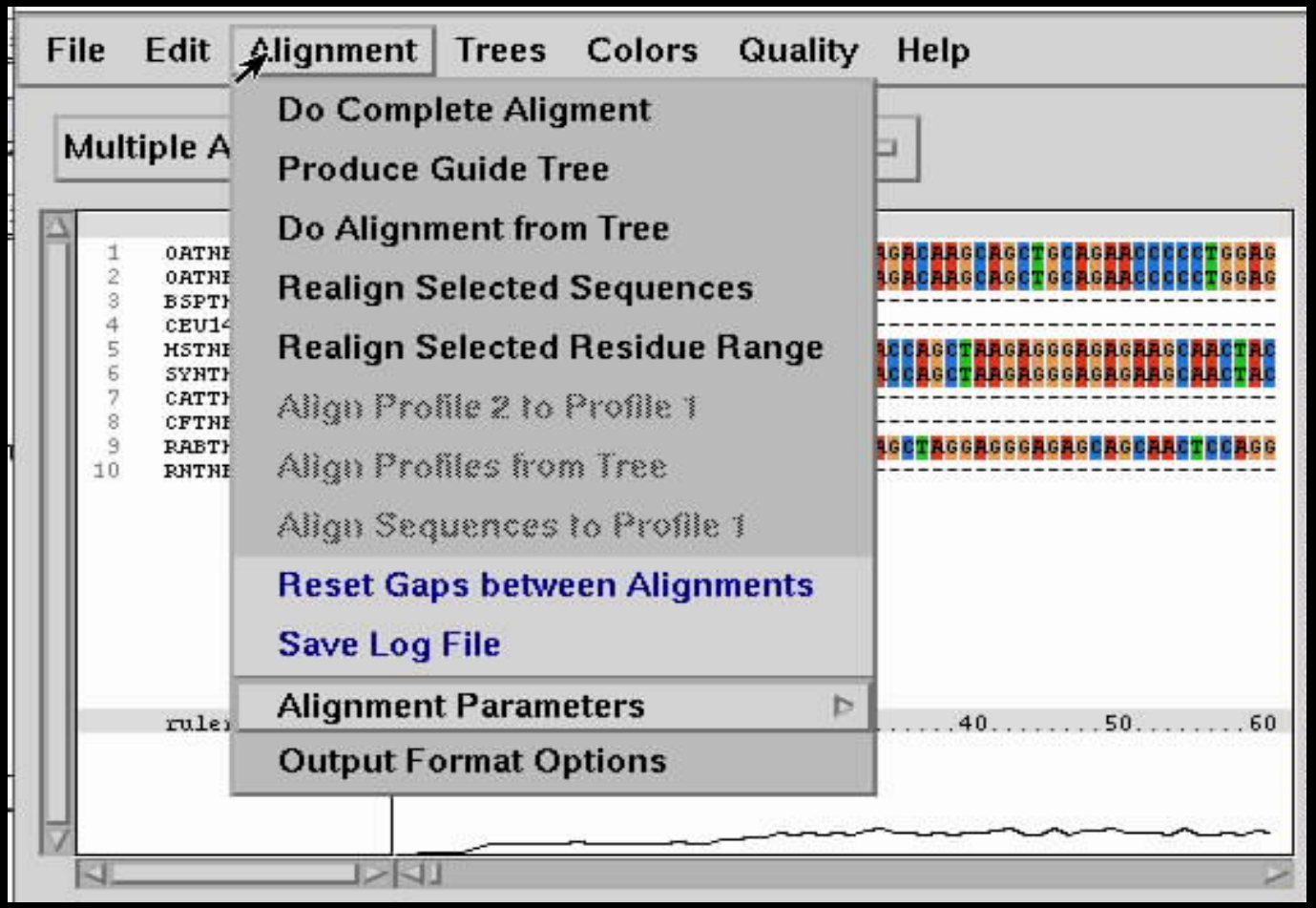
CLUSTALX


file_selection_popup
Filter
/users1/people/bfbecker/outputs/*
Directories
Files
/outputs/.
/outputs/..
/outputs/ameba
/outputs/others
/outputs/temp
coldna.par
colprint.par
colprot.par
databanks.txt
dotplot
eh_galectin.pep
eh_galectin.seg
eh_galectin.xnu
Selection
/users1/people/bfbecker/outputs/
OK
Filter
Cancel
Help

CLUSTALX


File
Edit
Alignment
Trees
Colors
Quality
Help
Multiple Alignment Mode
Font Size:
10
1 OATNFA1
2 OATNFAR
3 BSPTNFA
4 CEU14683
5 HSTNFR
6 SYNTNFTRP
7 CATTNFAA
8 CFTNFA
9 RABTNFM
10 RNTNFAA
ruler
.....430.....440.....450.....460.....470.....480.

CLUSTALX


File
Edit
Alignment
Trees
Colors
Quality
Help
Multiple A
Do Complete Aligment
Produce Guide Tree
Do Alignment from Tree
Realign Selected Sequences
Realign Selected Residue Range
Align Profile 2 to Profile 1
Align Profiles from Tree
Align Sequences to Profile 1
Reset Gaps between Alignments
Save Log File
Alignment Parameters
Output Format Options

PROS AND CONS OF PROGRESSIVE METHOD OF ALIGNMENT

PROS:

- Efficient enough to implement on a large scale for many (100s to 1000s) sequences.
- Progressive alignment services are commonly available on publicly accessible web servers, so users need not locally install the applications of interest.
- Most widely used method of multiple sequence alignment because of speed and accuracy.

CONS:

- Progressive alignments are not guaranteed to be globally optimal.
- The primary problem is that when errors are made at any stage in growing the MSA, these errors are then propagated through to the final result.
- Performance is also particularly bad when all of the sequences in the set are rather distantly related

REFERENCES

- Lipman DJ, Altschul SF, Kececioglu JD (1989). "A tool for multiple sequence alignment“.

- Hirosawa M, Totoki Y, Hoshida M, Ishikawa M (1995). "Comprehensive study on iterative algorithms of multiple sequence alignment“.

- Higgins DG, Sharp PM (1988). "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer".

THANK YOU