Dynamic Programming
(Local & Global Alignment)

Global alignment
the gap

The alignment is much better when one gap is introduced

Parameters for sequence alignment

Gap penalties

Opening: The cost to introduce a gap

Extension: The cost to extend a gap

Scoring systems

Every symbol pairing is assigned with a numerical value that is based on a „symbol comparison“ *or* „replacement“ table/matrix

Why gap penalties ?

The optimal alignment of two similar sequences usually

- **maximizes** the number of matches and
- **minimizes** the number of gaps.

Permitting the insertion of arbitrarily many gaps might lead to high scoring alignments of non-homologous sequences.

Penalizing gaps forces alignments to have relatively few gaps.

Gap penalties increase the quality of an alignment – non-homologous sequences are not aligned

Gap penalties

Linear gap penalty score:

$$\gamma(g) = - gd$$

Affine gap penalty score:

$$\gamma(g) = -d - (g - 1)\, e$$

$\gamma(g)$ = gap penalty score of a gap of length g

d = gap opening penalty

e = gap extension penalty

g = gap length

Scoring insertions and deletions

Total Score: 4

```
T A T G T G C G T A T A
  | | | |
  A T G T T A T A C
```

Total Score: 8 + (-3.2) = 4.8

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

Gap parameters:

d = 3 (gap opening)

e = 0.1 (gap extension)

g = 3 (gap length)

$\gamma(g) = -3 - (3 - 1)\ 0.1 = -3.2$

match = 1

mismatch = 0

Calculating alignments: Global vs. Local alignment

- For optimal GLOBAL alignment, we want best score in the final row or final column

  GLOBAL - best alignment of entirety of both sequences (possibly at expense of great local similarity)

- For optimal LOCAL alignment, we want best score anywhere in matrix

  LOCAL - best alignment of segments, without regard to rest of two sequences (at the expense of the overall score)

Important Points in Pairwise Sequence Alignment

Significance of Similarity

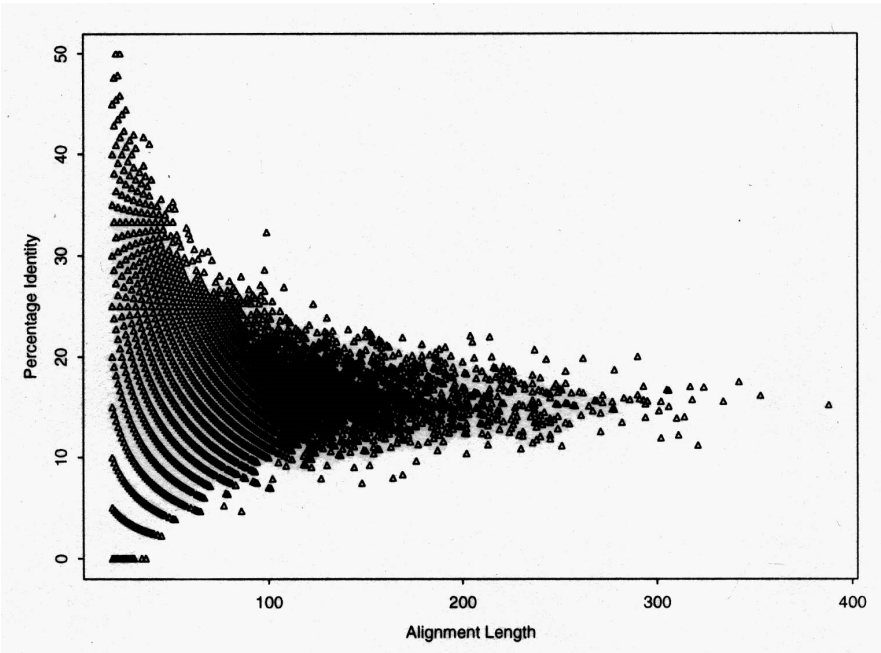
– Dependent on PID (Percent Identical Positions in Alignment)

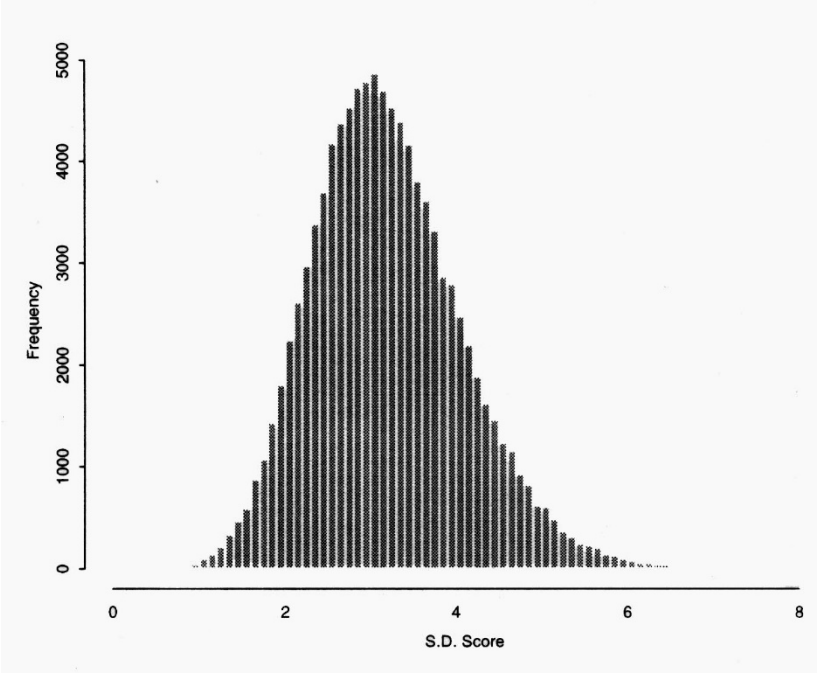
–Similarity/Disimilarity score

– Significance of score depend on length of alignment

–Significance Score (Z) whether score significant

–Expected Value (E), Chances that non-related sequence may have that score





Why we do multiple alignments?

- Multiple nucleotide or amino sequence alignment techniques are usually performed to fit one of the following scopes :
- In order to characterize protein families, identify shared regions of homology in a multiple sequence alignment; (this happens generally when a sequence search revealed homologies to several sequences)
- Determination of the consensus sequence of several aligned sequences.
- Help prediction of the secondary and tertiary structures of new sequences;
- Preliminary step in molecular evolution analysis using Phylogenetic methods for constructing phylogenetic trees

An example of Multiple Alignment

VTIS**C**TGSSSNIGAG-NHVK**W**YQ**QLPG**

VTIS**C**TGTSSNIGS--ITVN**W**YQ**QLPG**

LRLS**C**SSSGFIFSS--YAMY**W**VR**QAPG**

LSLT**C**TVSGTSFDD--YYST**W**VR**QPPG**

PEVT**C**VVVDVSHEDPQVKFN**W**YVDG--

ATLV**C**LISDFYPGA--VTVA**W**KADS--

AALG**C**LVKDYFPEP--VTVS**W**NSG---

VSLT**C**LVKGFYPSD--IAVE**W**WSNG--

Alignment of Multiple Sequences

Extending Dynamic Programming to more sequences

- –Dynamic programming can be extended for more than two
- –In practice it requires CPU and Memory (Murata et al 1985)
- – MSA, Limited only up to 8-10 sequences (1989)
- –DCA (Divide and Conquer; Stoye et al., 1997), 20-25 sequences
- –OMA (Optimal Multiple Alignment; Reinert et al., 2000)
- –COSA (Althaus et al., 2002)

Progressive or Tree or Hierarchical Methods (CLUSTAL-W)

- –Practical approach for multiple alignment
- –Compare all sequences pair wise
- –Perform cluster analysis
- –Generate a hierarchy for alignment
- –first aligning the most similar pair of sequences
- –Align alignment with next similar alignment or sequence

Alignment of Multiple Sequences

Iterative Alignment Techniques

- Deterministic (Non Stochastic) methods
  - They are similar to Progressive alignment
  - Rectify the mistake in alignment by iteration
  - Iterations are performed till no further improvement
  - AMPS (Barton & Sternberg; 1987)
  - PRRP (Gotoh, 1996), Most successful
  - Praline, IterAlign
- Stochastic Methods
  - SA (Simulated Annealing; 1994), alignment is randomly modified only acceptable alignment kept for further process. Process goes until converged
  - Genetic Algorithm alternate to SA (SAGA, Notredame & Higgins, 1996)
  - COFFEE extension of SAGA
  - Gibbs Sampler
  - Bayesian Based Algorithm (HMM; HMMER; SAM)
  - They are only suitable for refinement not for producing *ab initio* alignment. Good for profile generation. Very slow.

Alignment of Multiple Sequences

Progress in Commonly used Techniques (Progressive)

Clustal-W (1.8) (Thompson et al., 1994)

- Automatic substitution matrix
- Automatic gap penalty adjustment
- Delaying of distantly related sequences
- Portability and interface excellent

T-COFFEE (Notredame et al., 2000)

- Improvement in Clustal-W by iteration
- Pair-Wise alignment (Global + Local)
- Most accurate method but slow

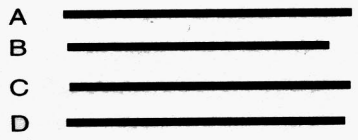
MAFFT (Katoh et al., 2002)

- Utilize the FFT for pair-wise alignment
- Fastest method
- Accuracy nearly equal to T-COFFEE

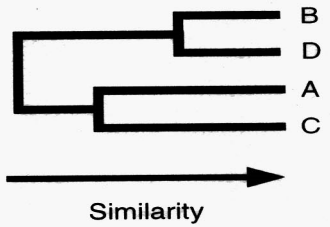
Steps in Multiple Alignment

(A) Pairwise Alignment

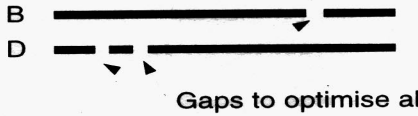
Example - 4 Sequences, A, B, C, D.



6 Pairwise Comparisons
then Cluster analysis

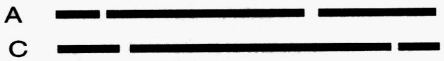


(B) Multiple alignment following the tree from A.



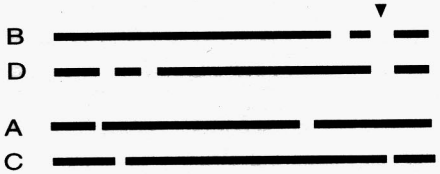
Gaps to optimise alignment

Align most similar pair.



Align next most similar pair.

New gap to optimise alignment of (BD) with (AC).



Align alignments - preserve gaps.

Multiple Alignment Method

- The steps are summarized as follows:
- Compare all sequences pairwise.
- Perform cluster analysis on the pairwise data
- Generate a hierarchy for alignment
  - Binary tree or a simple ordering
- First align the most similar pair of sequences
- Then the next most similar pair and so on.
- Once an alignment of two sequences has been made, then this is fixed.
- Thus for a set of sequences A, B, C, D having aligned
- A with C and B with D
- Alignment of A, B, C, D is obtained by comparing the alignments of A and C with that of B and D
  - using averaged scores at each aligned position.

ClustalW- for multiple alignment

- ClustaW is a multiple alignment program for DNA or proteins.
- Developed by Julie D. Thompson, Toby Gibson at EMBL/EBI
- ClustalW: Improving the sensitivity of multiple sequence alignment
  - sequence weighting
  - positions-specific gap penalties
  - weight matrix choice
  - Nucleic Acids Research, 22:4673-4680
- Manipulate existing alignments
- do profile analysis
- create phylogentic trees.
- Alignment can be done by 2 methods:
  - slow/accurate
  - fast/approximate

Running ClustalW

```
[~]%   clustalw

 ************************************************************
 ******** CLUSTAL W (1.7) Multiple Sequence Alignments  ********
 ************************************************************
```

1. Sequence Input From Disc
2. Multiple Alignments
3. Profile / Structure Alignments
4. Phylogenetic trees

S. Execute a system command
H. HELP
X. EXIT (leave program)

Your choice:

Using ClustalW

****** MULTIPLE ALIGNMENT MENU ******

1. Do complete multiple alignment now (Slow/Accurate)
2. Produce guide tree file only
3. Do alignment using old guide tree file

4. Toggle Slow/Fast pairwise alignments = SLOW

5. Pairwise alignment parameters
6. Multiple alignment parameters

7. Reset gaps between alignments? = OFF
8. Toggle screen display = ON
9. Output format options

S. Execute a system command
H. HELP
or press [RETURN] to go back to main menu

Your choice:

Output of ClustalW

```
CLUSTAL W (1.7) multiple sequence alignment


HSTNFR    GGGAAGAG---TTCCCCAGGGACCTCTCTCTAATCAGCCCTCTGGCCCAG------C
SYNTNFTRP GGGAAGAG---TTCCCCAGGGACCTCTCTCTAATCAGCCCTCTGGCCCAG------C
CFTNFA    ------------------------------------TGTCCAG------A
CATTNFAA  GGGAAGAG---CTCCCACATGGCCTGCAACTAATCAACCCTCTGCCCCAG------A
RABTNFM   AGGAGGAAGAGTCCCCAAACAACCTCCATCTAGTCAACCCTGTGGCCCAGATGGTCA
RNTNFAA   AGGAGGAGAAGTTCCCAAATGGGCTCCCTCTCATCAGTTCCATGGCCCAGACCCTCA
OATNFA1   GGGAAGAGCAGTCCCCAGCTGGCCCCTCCTTCAACAGGCCTCTGGTTCAG------A
OATNFAR   GGGAAGAGCAGTCCCCAGCTGGCCCCTCCTTCAACAGGCCTCTGGTTCAG------A
BSPTNFA   GGGAAGAGCAGTCCCCAGGTGGCCCCTCCATCAACAGCCCTCTGGTTCAA------A
CEU14683  GGGAAGAGCAATCCCCAACTGGCCTCTCCATCAACAGCCCTCTGGTTCAG------A
                                                          **
```

ClustalW options

Your choice: 5

********* PAIRWISE ALIGNMENT PARAMETERS *********

Slow/Accurate alignments:

1. Gap Open Penalty :15.00
2. Gap Extension Penalty :6.66
3. Protein weight matrix :BLOSUM30
4. DNA weight matrix :IUB

Fast/Approximate alignments:

5. Gap penalty :5
6. K-tuple (word) size :2
7. No. of top diagonals :4
8. Window size :4

9. Toggle Slow/Fast pairwise alignments = SLOW

H. HELP

Enter number (or [RETURN] to exit):

ClustalW options

Your choice: 6

********* MULTIPLE ALIGNMENT PARAMETERS *********

1. Gap Opening Penalty :15.00
2. Gap Extension Penalty :6.66
3. Delay divergent sequences :40 %

4. DNA Transitions Weight :0.50

5. Protein weight matrix :BLOSUM series
6. DNA weight matrix :IUB
7. Use negative matrix :OFF

8. Protein Gap Parameters

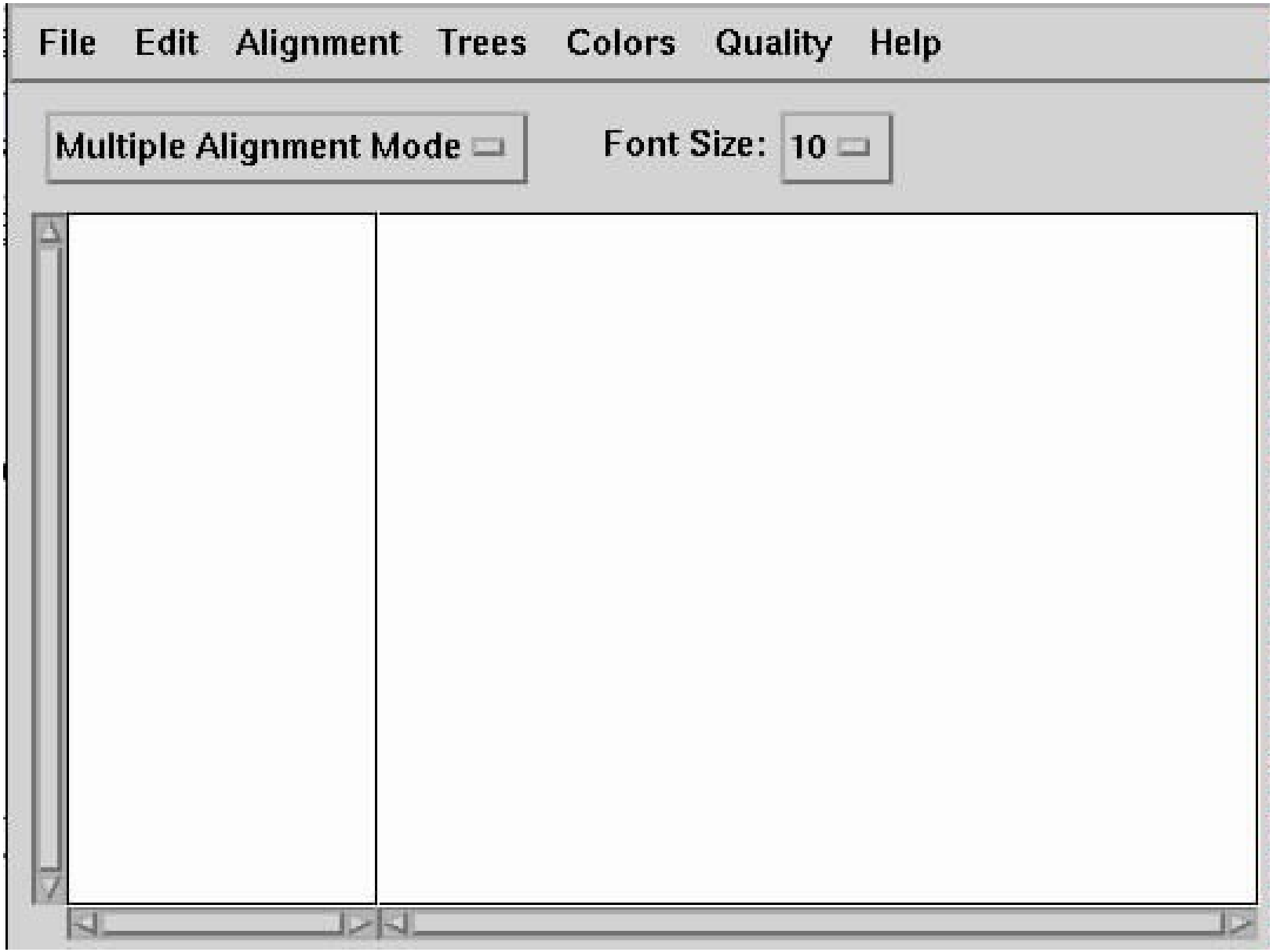
H. HELP

Enter number (or [RETURN] to exit):

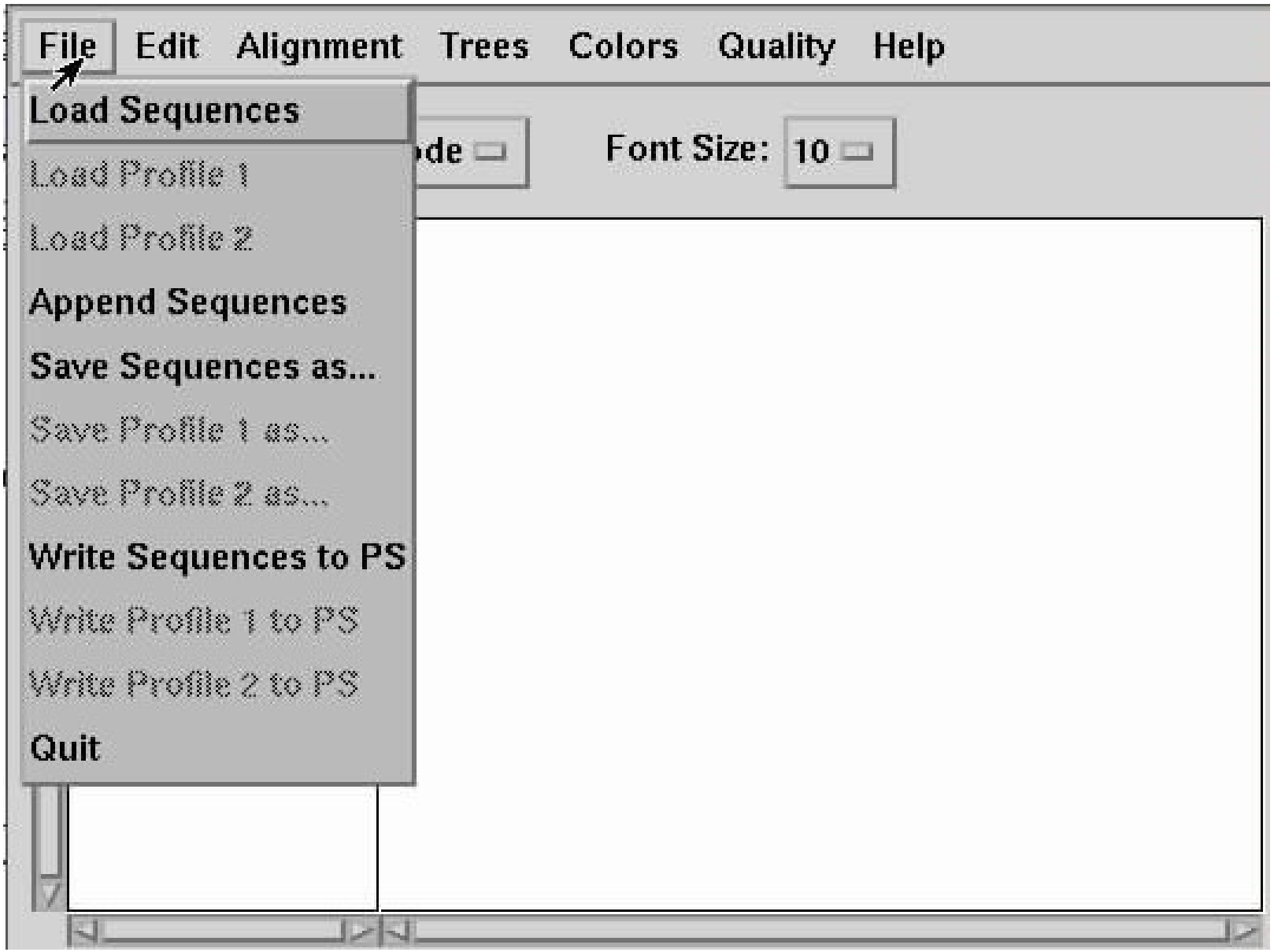
ClustalX - Multiple Sequence Alignment Program

- 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 Information (Bldg 38A, NIH 8600 Rockville Pike,Bethesda, MD 20894) as part of their NCBI SOFTWARE DEVELOPEMENT TOOLKIT.

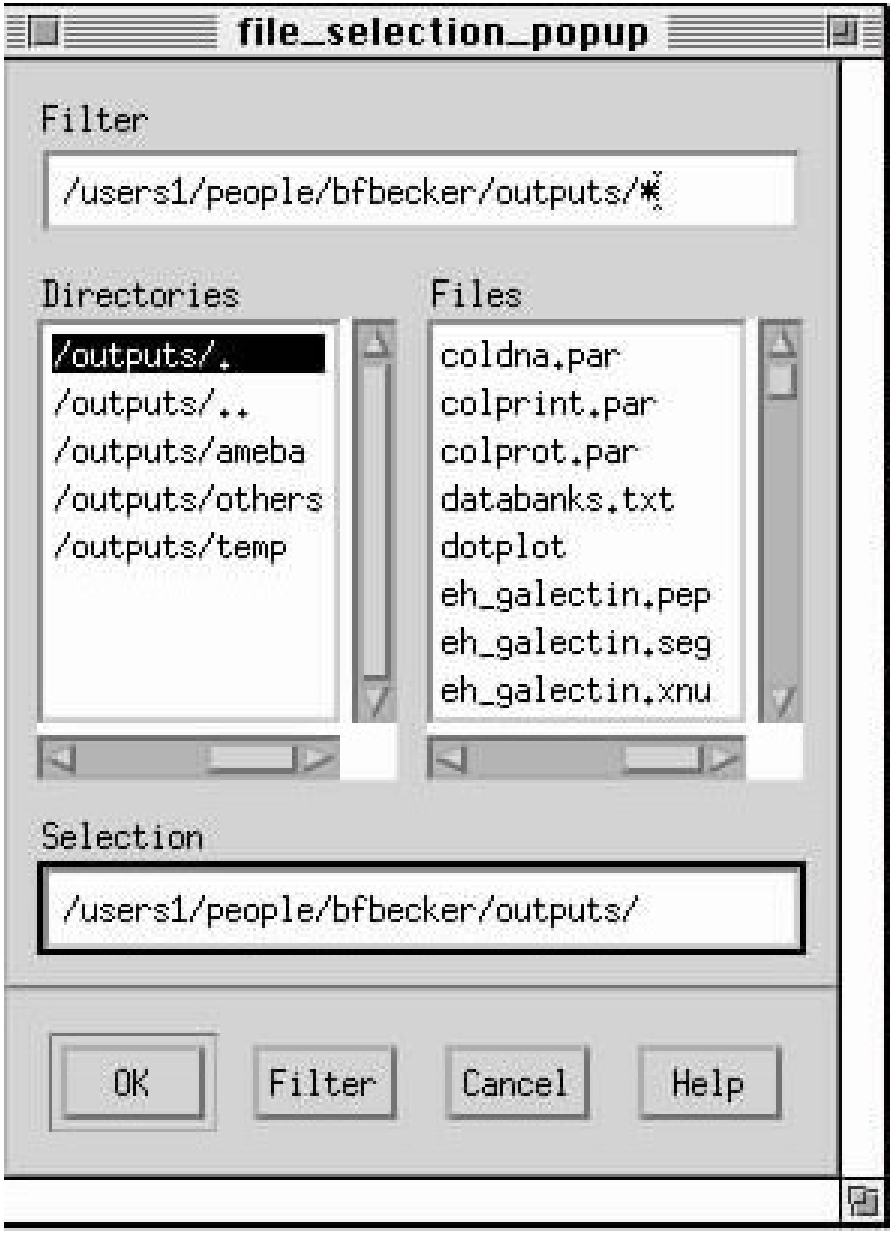
ClustalX


File
Edit
Alignment
Trees
Colors
Quality
Help
Multiple Alignment Mode
Font Size:
10

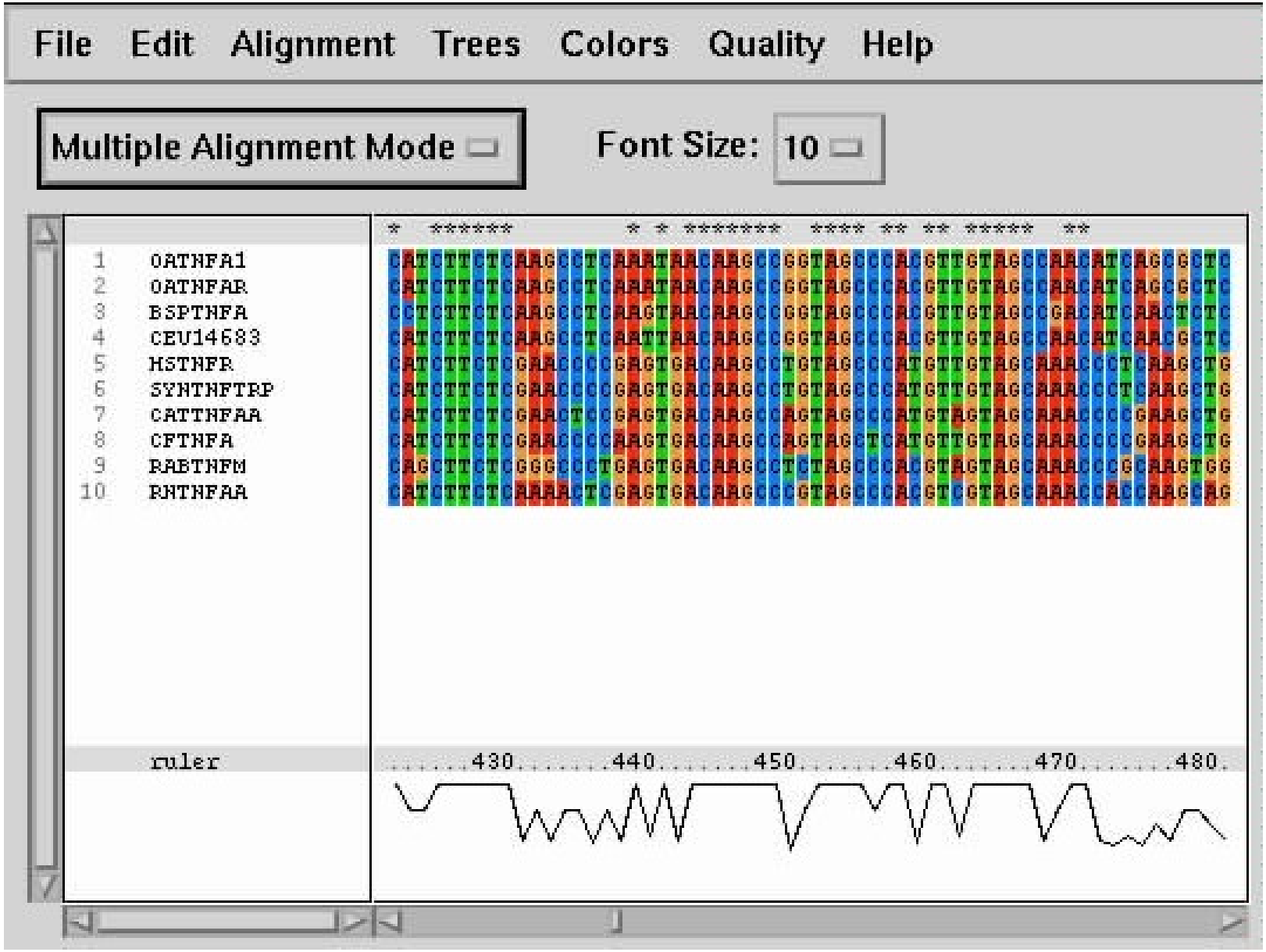
ClustalX


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

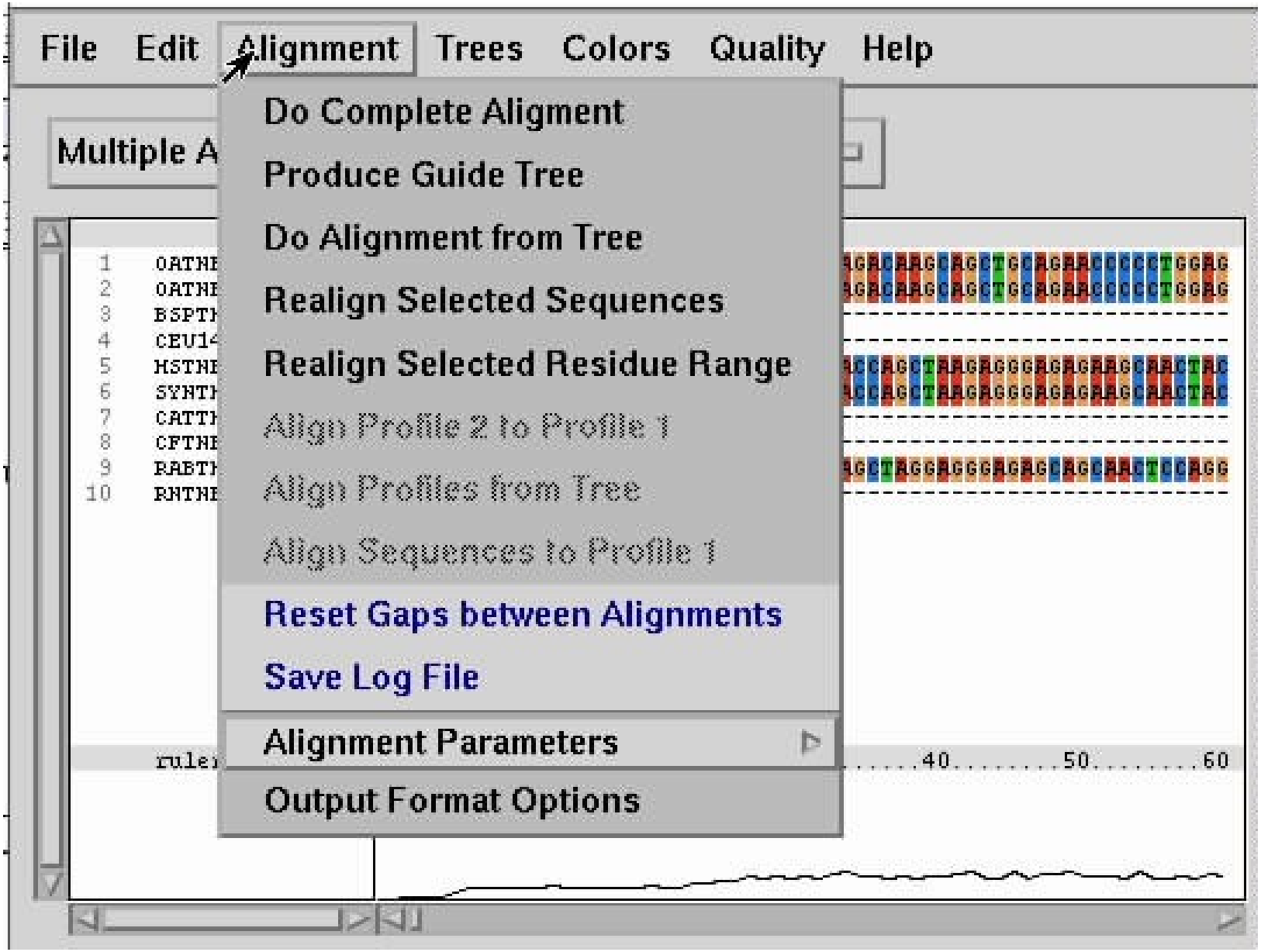
ClustalX


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

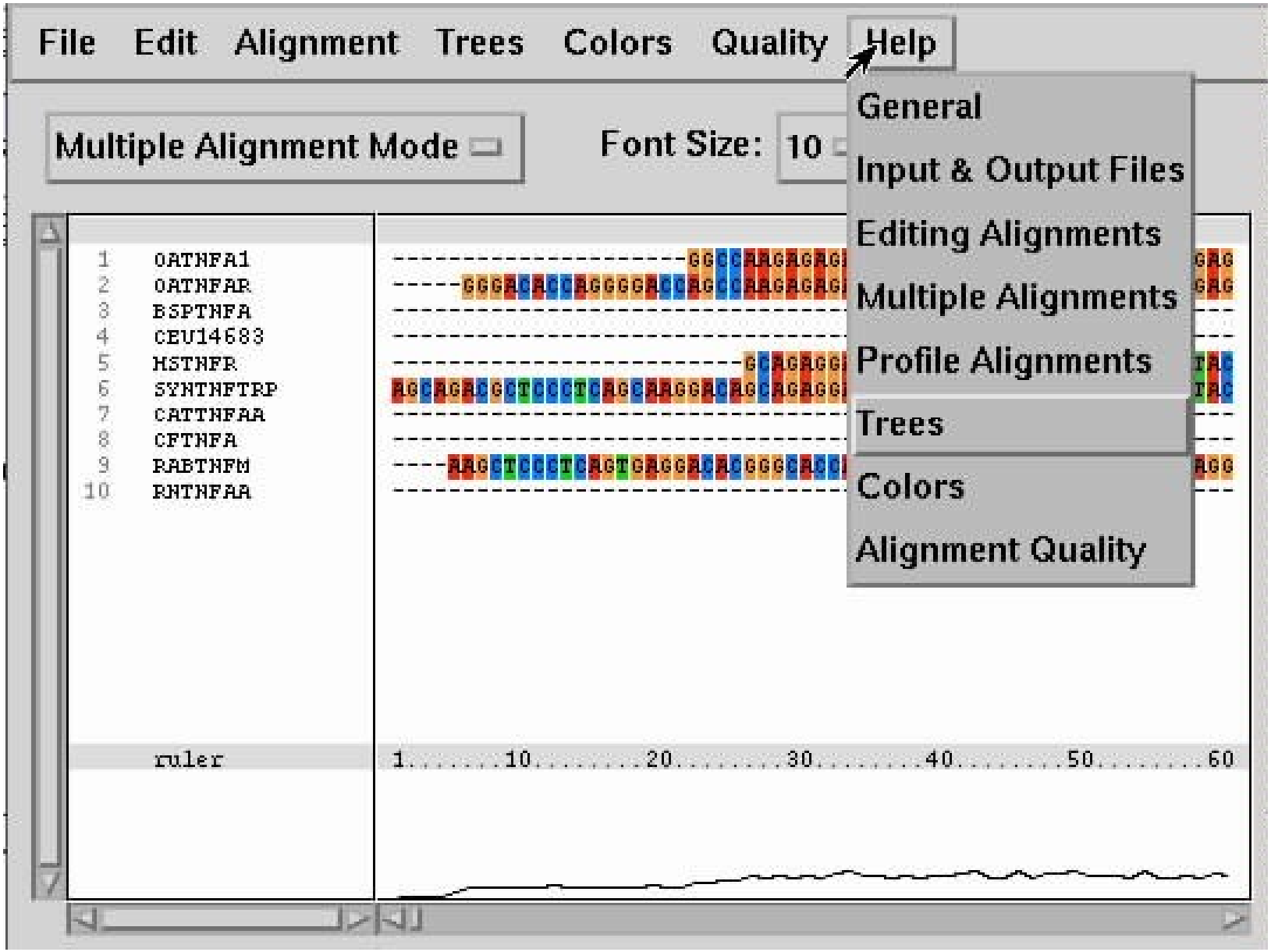
ClustalX


File Edit Alignment Trees Colors Quality Help
Multiple Alignment Mode
Font Size: 10
1 OATNFA1
2 OATNFAR
3 BSPTNFA
4 CEU14683
5 MSTNFR
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

ClustalX


File
Edit
Alignment
Trees
Colors
Quality
Help
General
Input & Output Files
Editing Alignments
Multiple Alignments
Profile Alignments
Trees
Colors
Alignment Quality
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
1.....10.....20.....30.....40.....50.....60