Applied Bioinformatics
February 21, 2011

Multiple Sequence Alignment

Zhongming Zhao, PhD
Email: zhongming.zhao@vanderbilt.edu
http://bioinfo.mc.vanderbilt.edu/

The process of aligning sequences is a game involving playing off gaps and mismatches

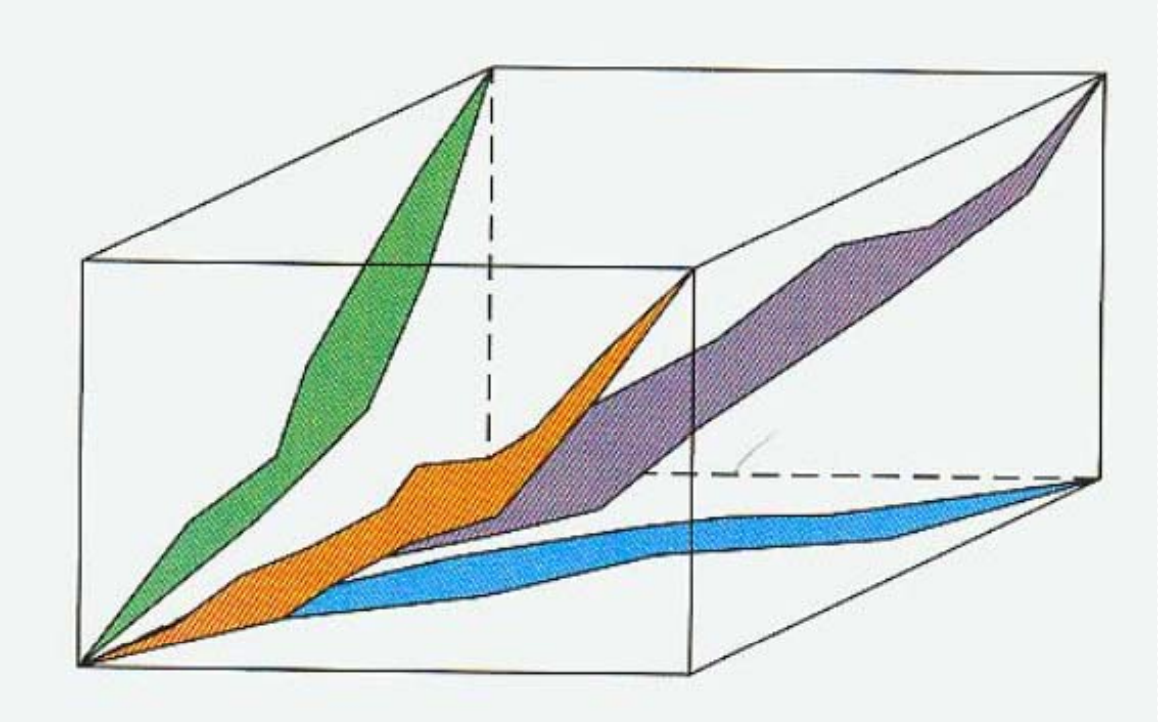
Ways of Aligning Multiple Sequences

- By hand based on knowledge/experience
  - Specific sorts of columns in alignment, such as highly conserved residues or buried hydrophobic residues
  - The influence of secondary and tertiary structure, such as the alteration of hydrophobic and hydrophilic columns in exposed beta sheet
  - Expected patterns of insertions and deletions
  - Tedious, error-prone
- Automated
  - Assign a score to find the “best” multiple alignments
  - Uncertainty of the “true” alignment
- Combination
  - Errors may come from both manual or computational approach
  - The structure and evolutionary factors should be considered

MSA

- The principle of dynamic programming in pairwise alignment can be extended to multiple sequences
- Unfortunately, the time required grows exponentially with the number of sequences and sequence lengths, this turns out to be impractical.
- Algorithms in use are heuristic and most are progressive/hierarchical

Multidimensional Dynamic Programming



An optimal alignment is found by MSA for three sequences.
From David Mount text book Bioinformatics

Multidimensional Dynamic Programming

$\alpha_{i_1,i_2,\ldots,i_N}$: the maximum score of an alignment up to the subsequences ending with $x^1_{i_1}, x^2_{i_2}, \ldots, x^N_{i_N}$. The dynamic programming algorithm is

$$\alpha_{i_1,i_2,\ldots,i_N} = \max \begin{cases} \alpha_{i_1-1,i_2-1,\ldots,i_N-1} + S(x^1_{i_1}, x^2_{i_2}, \ldots, x^N_{i_N}), \\ \alpha_{i_1,i_2-1,\ldots,i_N-1} + S(-, x^2_{i_2}, \ldots, x^N_{i_N}), \\ \alpha_{i_1-1,i_2,i_3-1,\ldots,i_N-1} + S(x^1_{i_1}, -, \ldots, x^N_{i_N}), \\ \vdots \\ \alpha_{i_1-1,i_2-1,\ldots,i_N} + S(x^1_{i_1}, x^2_{i_2}, \ldots, -), \\ \alpha_{i_1,i_2,i_3-1,\ldots,i_N-1} + S(-, -, \ldots, x^N_{i_N}), \\ \vdots \\ \alpha_{i_1,i_2-1,\ldots,i_{N-1}-1,i_N} + S(-, x^2_{i_2}, \ldots, -), \\ \vdots \end{cases}$$

Where all combinations of gaps appear except the one where all residues are replaced by gaps. Gap penalty, initialization, termination, and traceback follow the pairwise dynamic programming algorithm.

Multiple Alignment Programs

- Biopat (first method ever)
- MSA (Lipman et al 1989)
- MULTAL (Taylor 1987)
- DIALIGN (Morgenstern 1996)
- PRRP (Gotoh 1996)
- PILEUP (GCG package)
- Clustal W/W2/X (Thompson Higgins Gibson 1994)
- Praline (Heringa 1999)
- T-COFFEE (Poirot et al. 2003)
- HMMER (Eddy 1998) [Hidden Markov Models]
- SAGA (Notredame 1996) [Genetic algorithms]
- MEME, MULTIPIPMAKER, et al.

- http://pbil.univ-lyon1.fr/alignment.html

Approaches to MSA

- Progressive alignment methods
- Iterative refinement methods

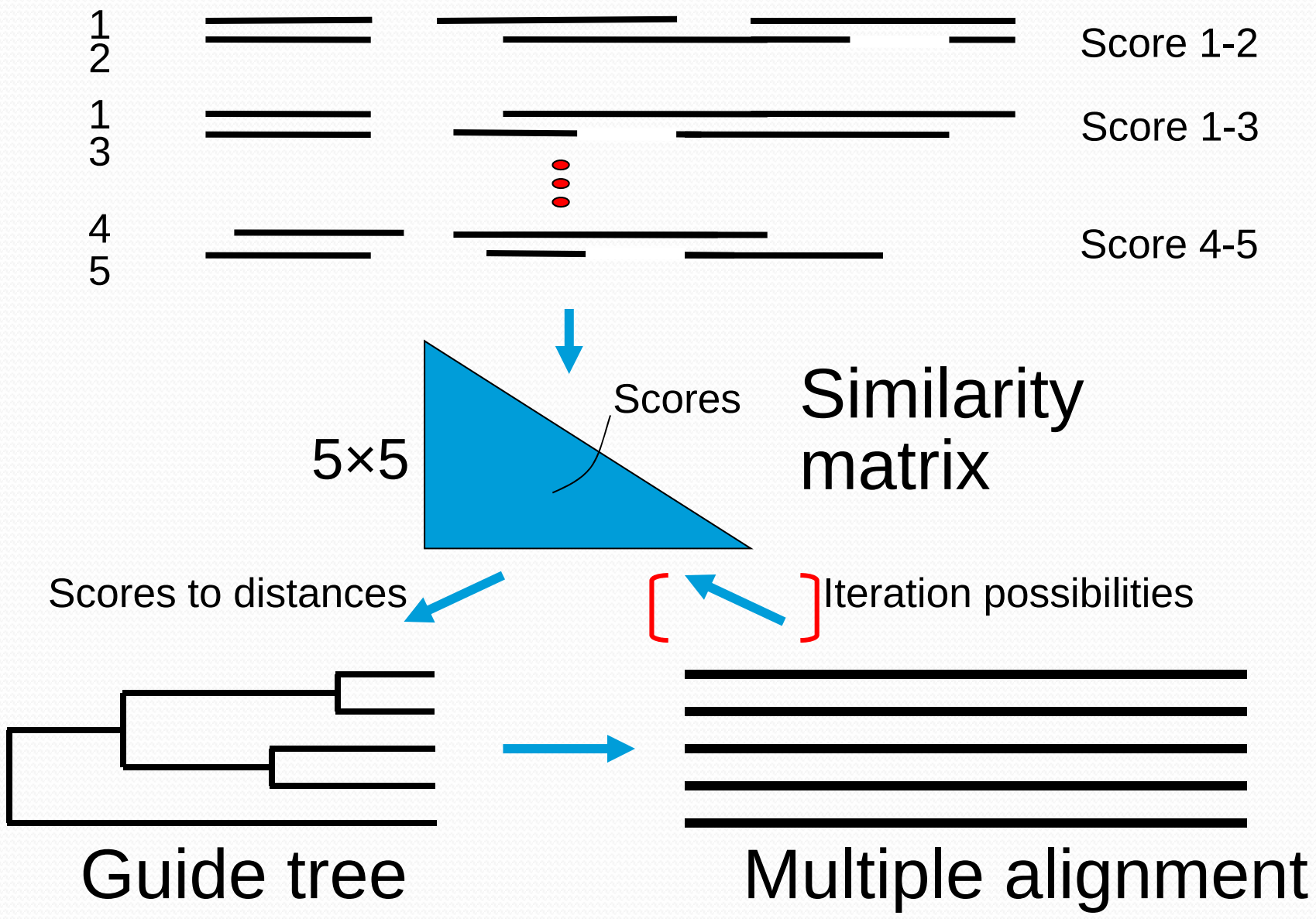
Progressive Alignment Methods

- This approach is the most commonly used in MSA.
  - Two sequences are chosen and aligned by standard pairwise alignment; this alignment is fixed.
  - A third sequence is chosen and aligned to the first alignment
  - This process is iterated until all sequences have been aligned
- This approach was applied in a number of algorithms, which differ in
  - How to choose the order to do the alignment
  - Whether the progression involves only alignment of sequences to a single growing alignment or whether subfamilies are built up on a tree structure and, at certain points, alignments are aligned to alignments
  - Procedure used to align and score sequences or alignments against existing alignments.

Progressive Alignment Methods

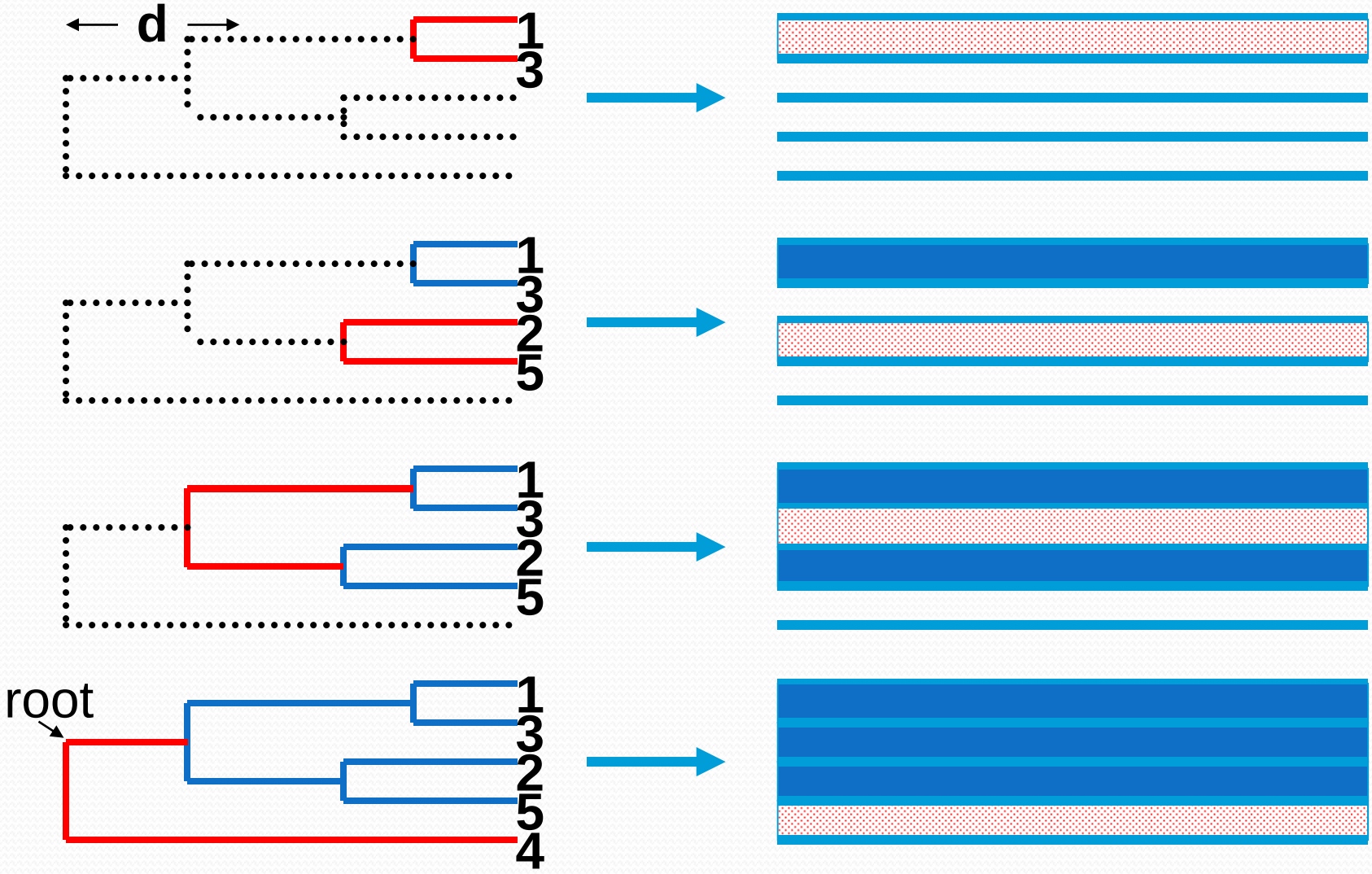
- Advantages
  - Fast
  - Efficient
  - The resulting alignments are reasonable in may cases
- Disadvantages
  - Heuristic
  - Accuracy is very important
  - Errors are propagated into the progressive steps

Progressive Multiple Alignment General Principles



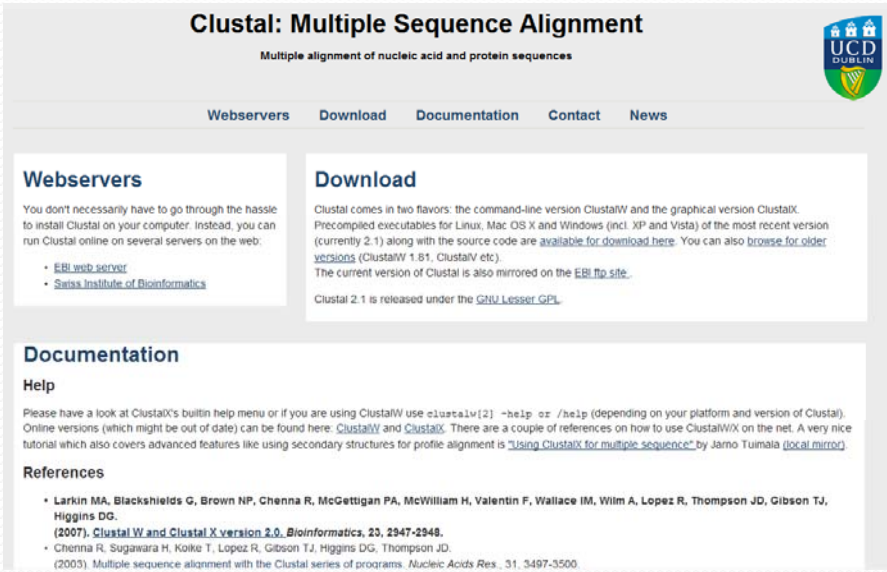
General Progressive Multiple Alignment Technique

(follow generated tree)



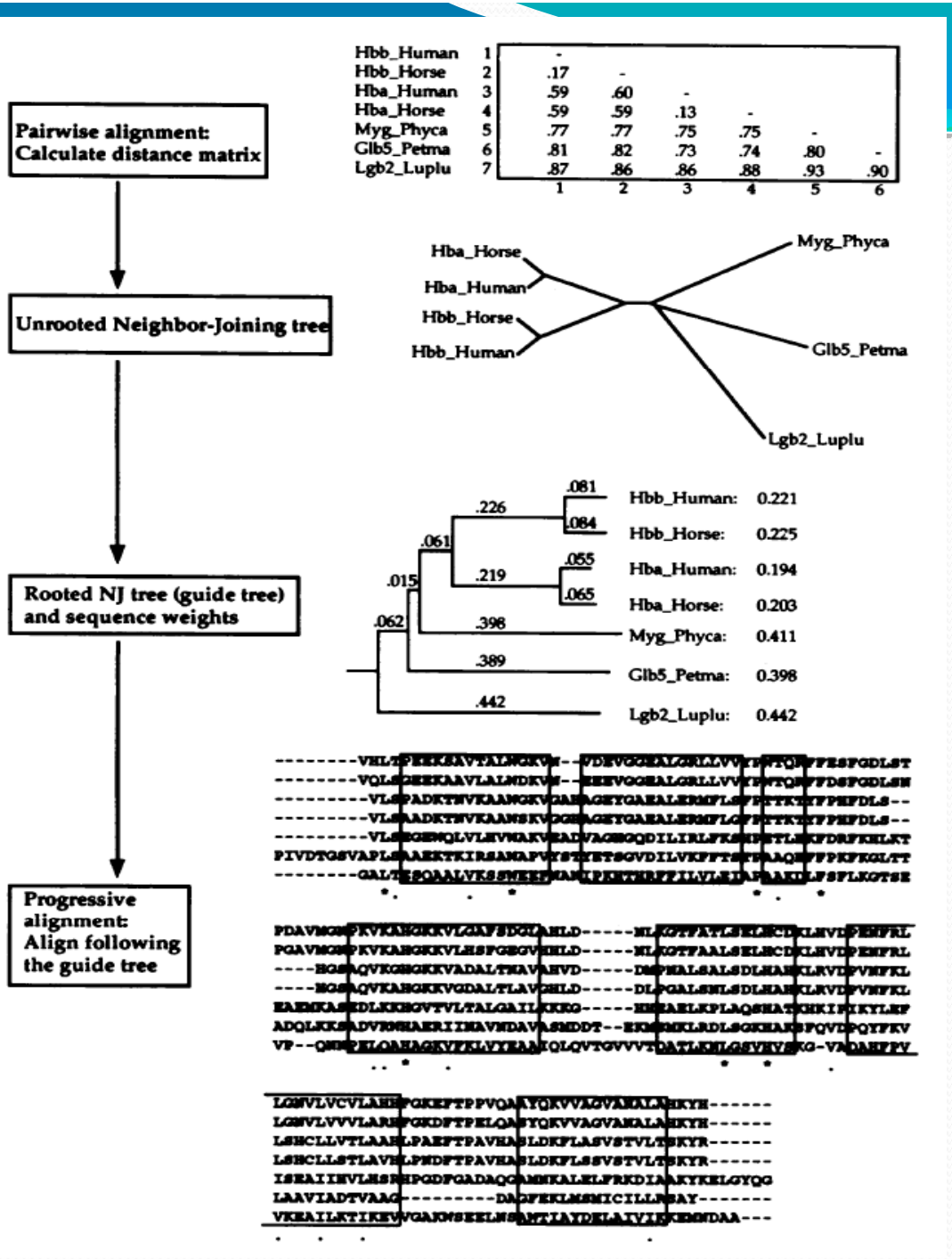
Clustal W

- The widely used profile-based progressive multiple alignment (Thompson, Higgins, and Gibson 1994, Nucl. Acids Res, authors from EMBL-Heidelberg).
- Succeeded from Clustal V
- W means weighting
- It is carefully tuned use of profile alignment methods.
- Clustal X provides the graphic interface utility.
- http://www.clustal.org/

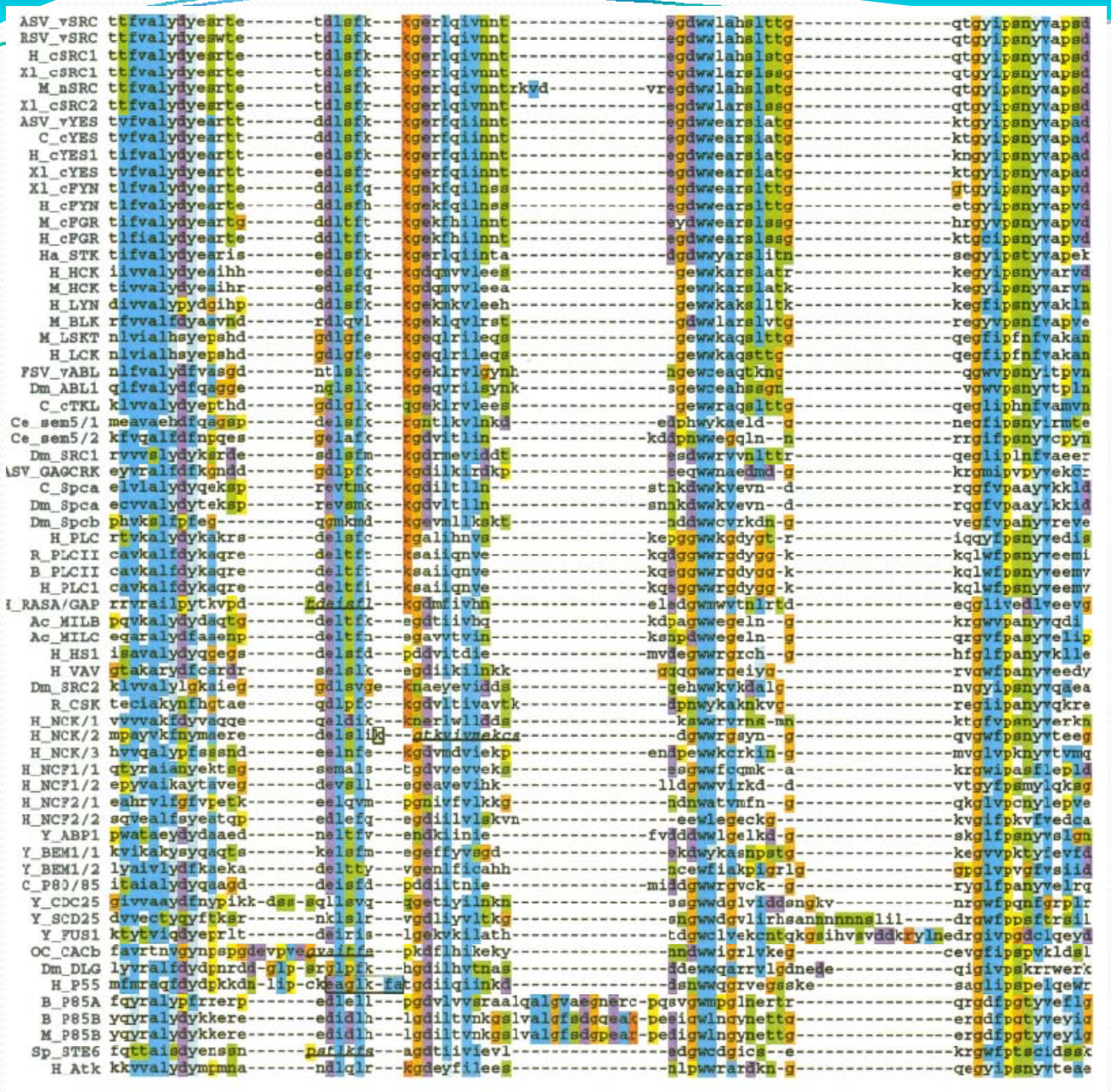


Clustal W

- Algorithm
  - Construct a distance matrix of all N(N-1)/2 pairs by pairwise dynamic programming alignment followed by approximate conversion of similarity scores to evolutionary distances using the model of Kimura (1983)
  - Construct a guide tree by a Neighbor-Joining method (Saitou and Nei 1987) from the distance matrix
  - Progressively align at nodes in order of decreasing similarity, as in the guide tree, using sequence-sequence, sequence-profile, and profile-profile alignment.
- Many heuristic improvements make the Clustal W an accurate algorithm.
  - Sequence weighting
  - Gap and gap extension
  - Divergence of sequences


Pairwise alignment: Calculate distance matrix
Unrooted Neighbor-Joining tree
Rooted NJ tree (guide tree) and sequence weights
Progressive alignment: Align following the guide tree
Hbb_Human 1
Hbb_Horse 2 .17
Hba_Human 3 .59 .60
Hba_Horse 4 .59 .59 .13
Myg_Phyca 5 .77 .77 .75 .75
Glb5_Petma 6 .81 .82 .73 .74 .80
Lgb2_Luplu 7 .87 .86 .86 .88 .93 .90
1 2 3 4 5 6
Hba_Horse
Hba_Human
Hbb_Horse
Hbb_Human
Myg_Phyca
Glb5_Petma
Lgb2_Luplu
.081
.226
.084
.061
.055
.219
.015
.065
.062
.398
.389
.442
Hbb_Human: 0.221
Hbb_Horse: 0.225
Hba_Human: 0.194
Hba_Horse: 0.203
Myg_Phyca: 0.411
Glb5_Petma: 0.398
Lgb2_Luplu: 0.442

Clustal W Alignment of a Set of SH3 Domains



SH3 domains have a minimum similarity below 12% identity, poorly aligned by other programs, which did not generate the correct blocks for 2nd structure.

hydrophobic = blue

hydrophobic tendency = light blue

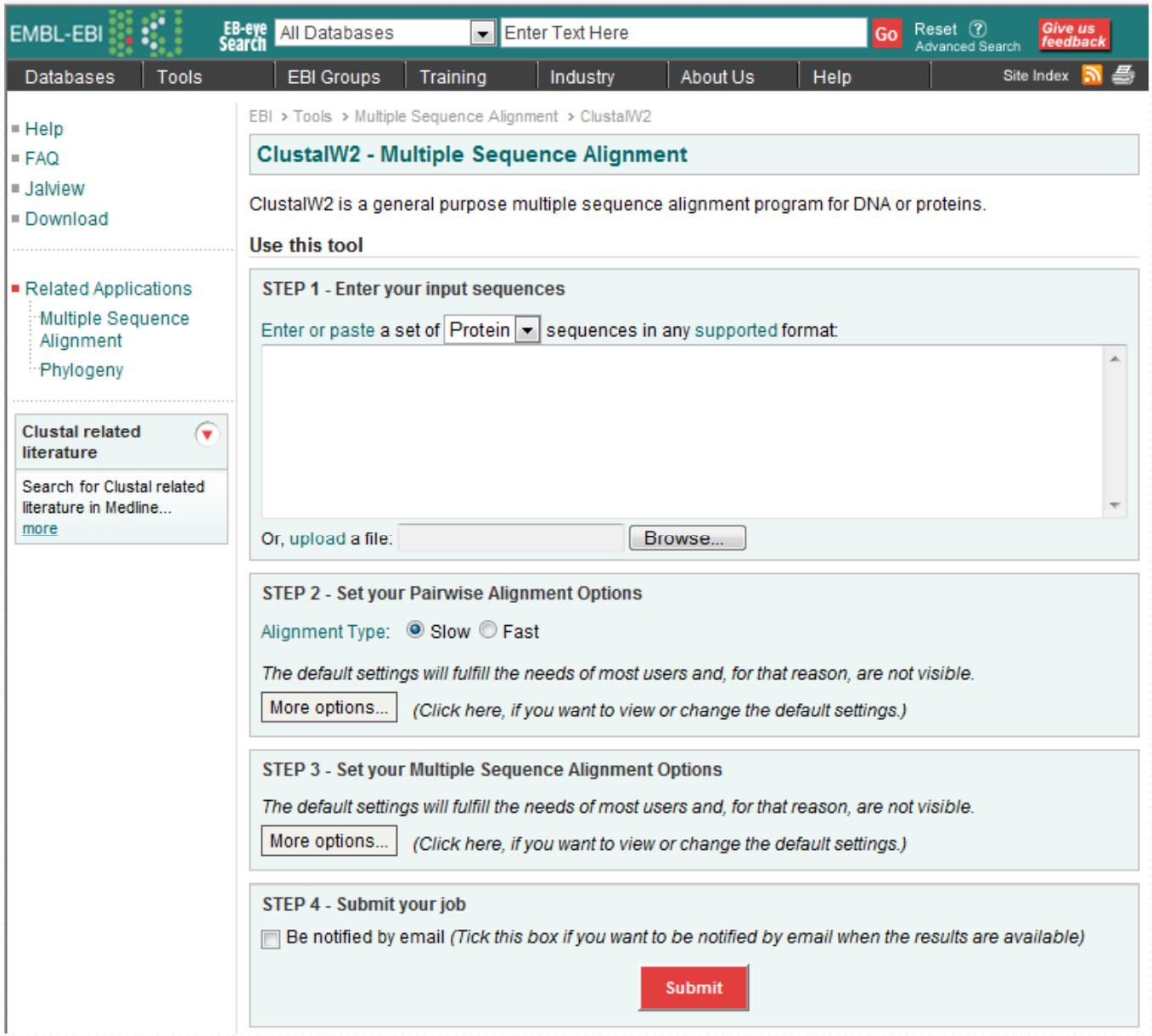
basic = red

acidic = purple

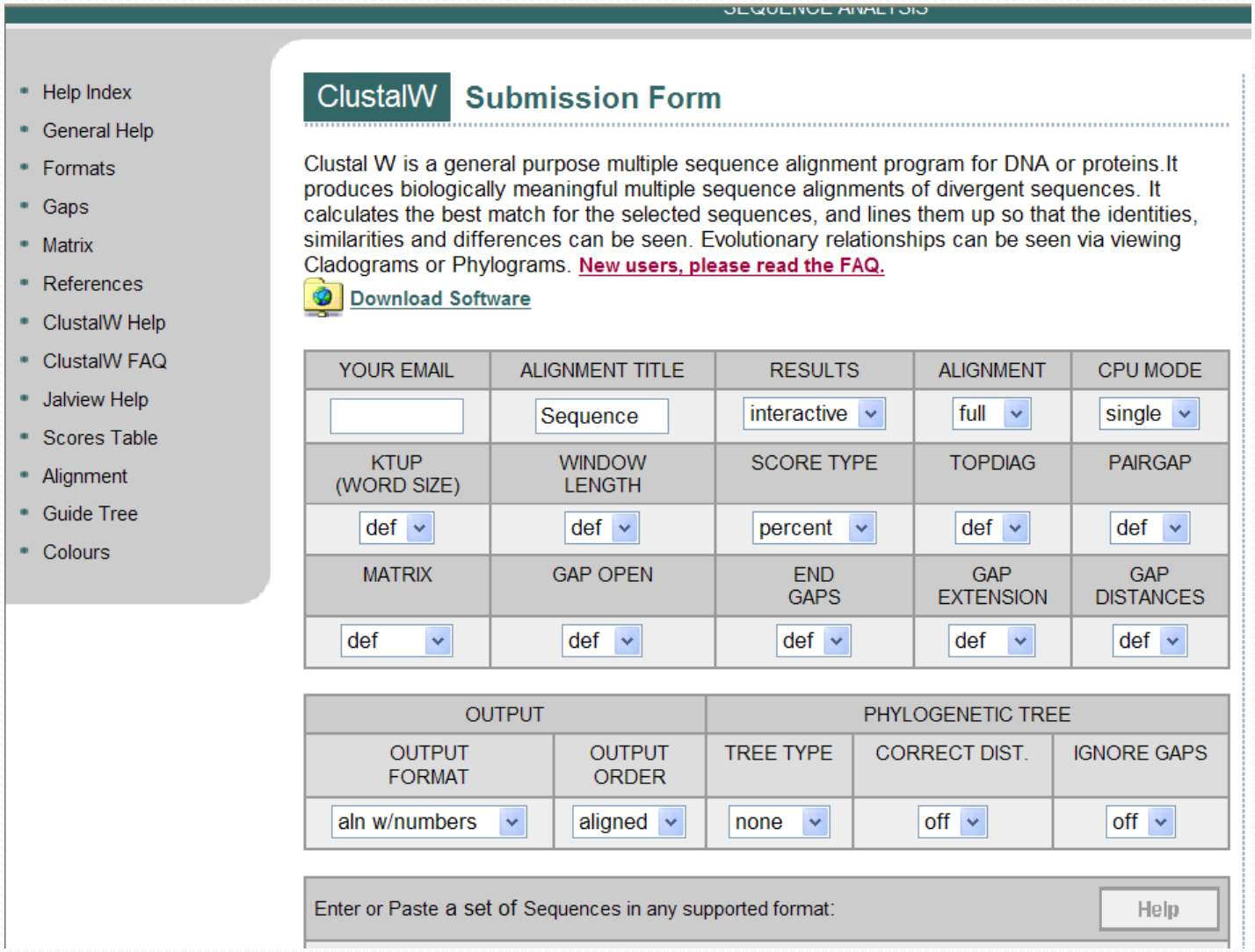
hydrophilic = green

unconserved = white

http://www.ebi.ac.uk/clustalw/



Back to 2005: http://www.ebi.ac.uk/clustalw/



SEQUENCE ANALYSIS

- Help Index
- General Help
- Formats
- Gaps
- Matrix
- References
- ClustalW Help
- ClustalW FAQ
- Jalview Help
- Scores Table
- Alignment
- Guide Tree
- Colours

ClustalW Submission Form

Clustal W is a general purpose multiple sequence alignment program for DNA or proteins.It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms. **New users, please read the FAQ.**

Download Software

YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	CPU MODE
	Sequence	interactive	full	single
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	def	def	def

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers	aligned	none	off	off

Enter or Paste a set of Sequences in any supported format: Help

Clustal X

http://www.clustal.org/download/current/

Clustal Format

CLUSTAL W (1.82) multiple sequence alignment

"*" means that the residues or nucleotides in that column are identical in all sequences in the alignment.

":" means that conserved substitutions have been observed.

"." means that semi-conserved substitutions are observed.

```
CLUSTAL W (1.82) multiple sequence alignment


FOSB_MOUSE      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
FOSB_HUMAN      MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA 60
                ************************************************************

FOSB_MOUSE      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS 120
FOSB_HUMAN      ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPVVDPYDMPGTSYSTPGMSGYSSGGASGS 120
                ********************************.***************:*.**:******
```

Multiple Alignment Strategies

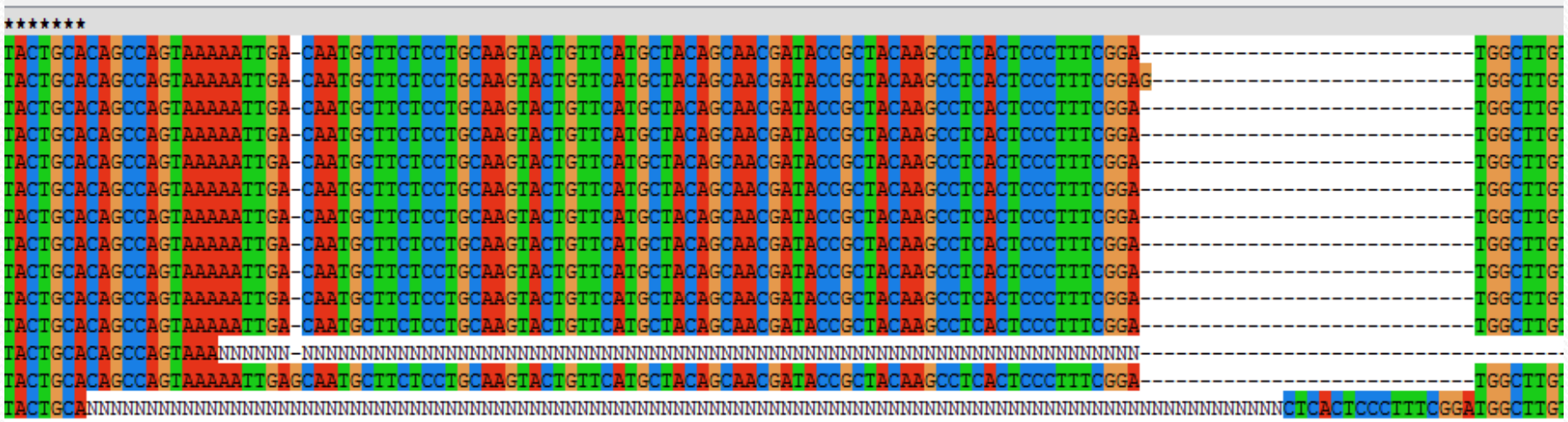
- Align pairs of sequences using an optimal method
- Choose representative sequences to align carefully
- Choose sequences of comparable lengths
- Progressive alignment programs such as Clustal X for multiple alignment
- Progressive alignment programs may be combined
- Review alignment by eye and edit

Multiple Alignments and Phylogenetic Trees

- You can make a more accurate multiple sequence alignment if you know the tree already
- A good multiple sequence alignment is an important starting point for drawing a tree
- The process of constructing a multiple alignment (unlike pairwise) needs to take account of phylogenetic relationships

Editing a Multiple Sequence Alignment

- It is NOT fraud to edit a multiple sequence alignment
- Incorporate additional knowledge if possible
- Alignment editors help to keep the data organized and help to prevent unwanted mistakes



An Example

- Align 14 SARS “complete” genome sequences
- Cut first 20000 bp and aligned them
- Examine the alignments (e.g. 8528), need to adjust by hand!