Multiple Sequence Alignment

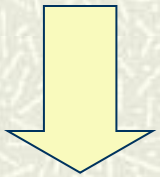
Multiple Sequence Alignment

- Collection of three or more amino acid (or nucleic acid) sequences partially or completely aligned.

- Aligned residues tend to occupy corresponding positions in the 3-D structure of each aligned protein.

General steps to multiple alignment.

Create Alignment



Edit the alignment to ensure that regions of functional or structural similarity are preserved

USED FOR:

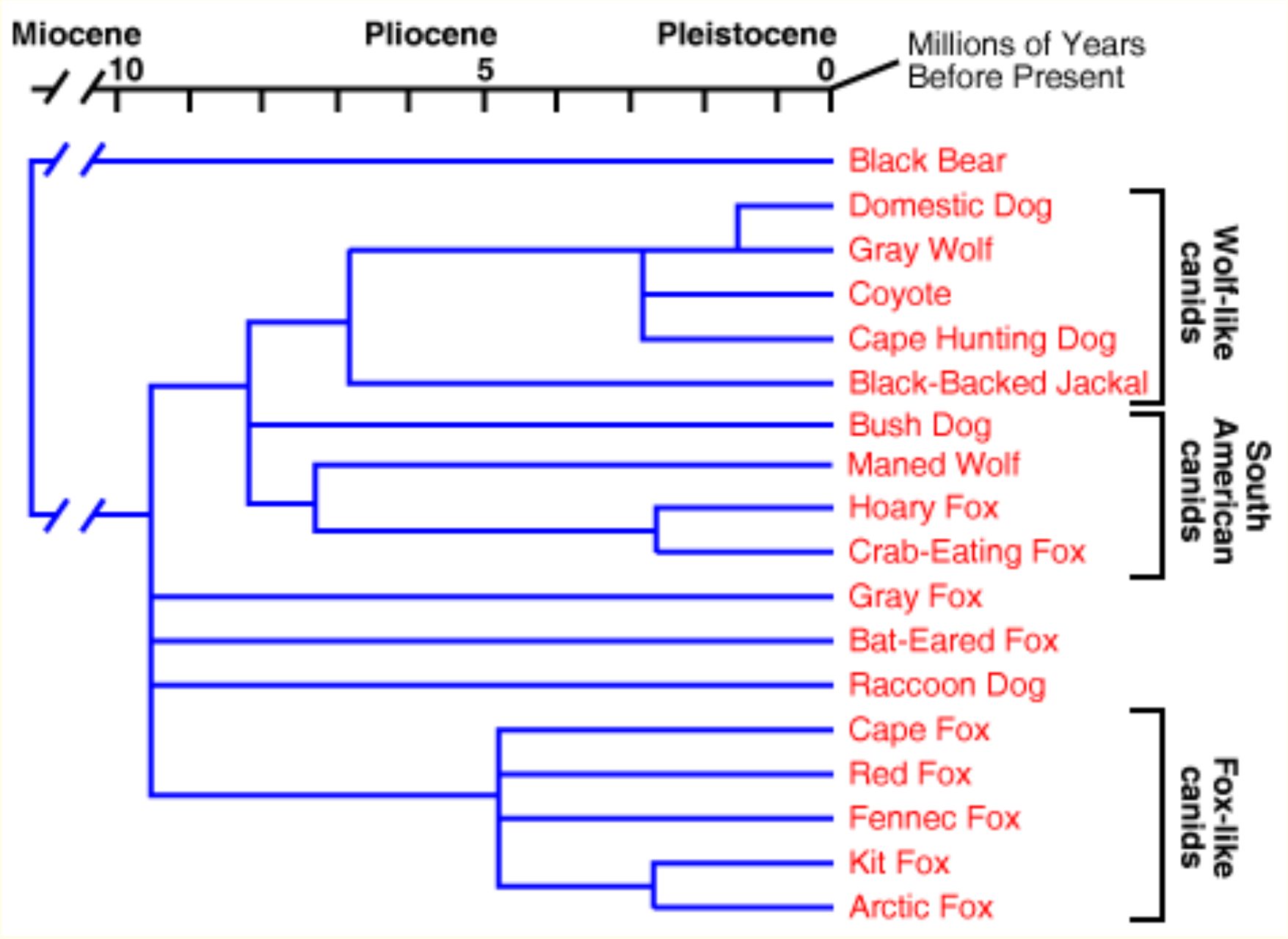
Phylogenetic Analysis	Structure Analysis	Find conserved motifs to deduce function	Design of PCR primers

Practical use of MSA

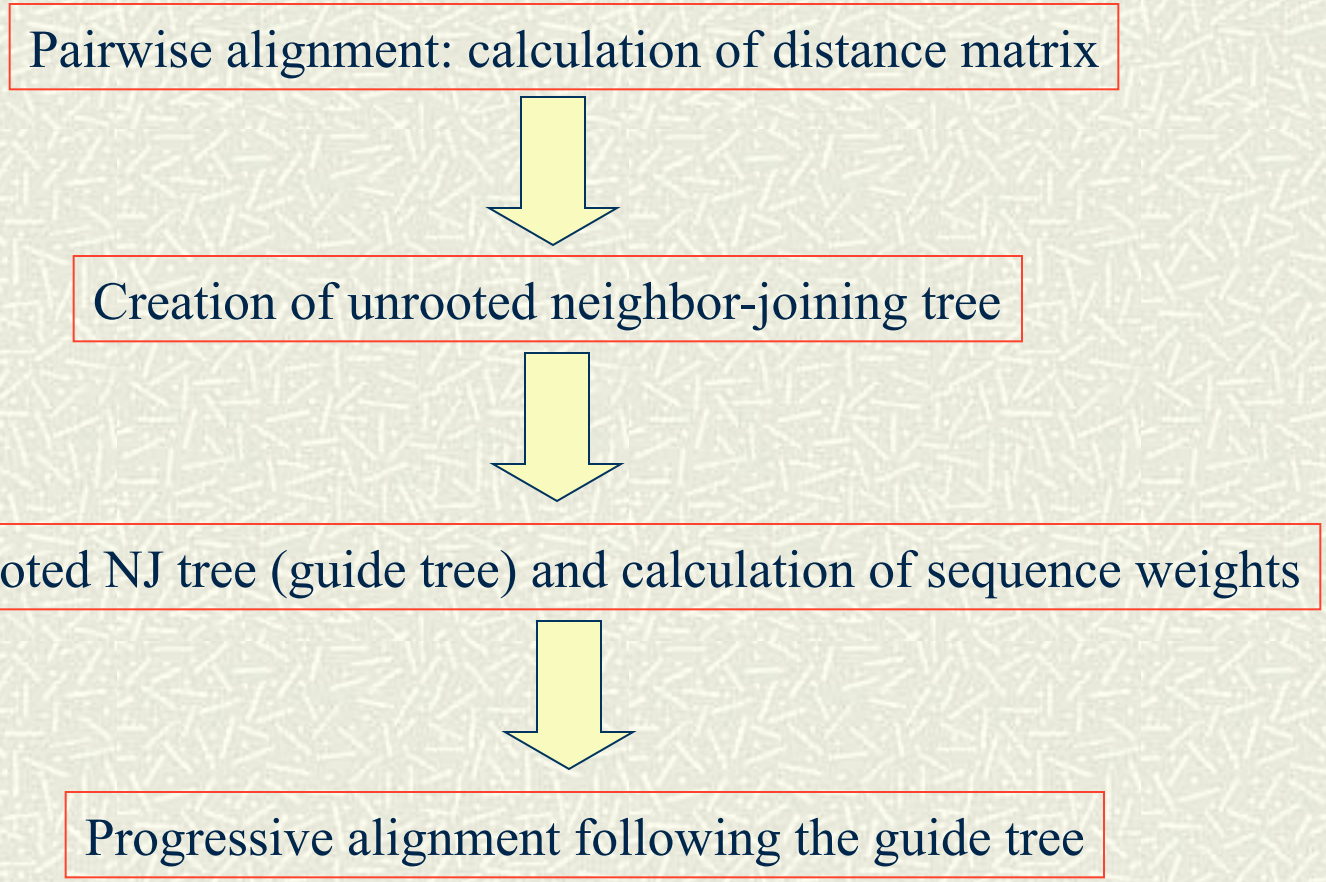
- Helps to place protein into a group of related proteins. It will provide insight into function, structure and evolution.
- Helps to detect homologs
- Identifies sequencing errors
- Identifies important regulatory regions in the promoters of genes.

Clustal W (Thompson et al., 1994)

- CLUSTAL=Cluster alignment
- The underlying concept is that groups of sequences are phylogenetically related. If they can be aligned, then one can construct a phylogenetic tree.
- Phylogenetic tree-a tree showing the evolutionary relationships among various biological species or other entities that are believed to have a common ancestor.

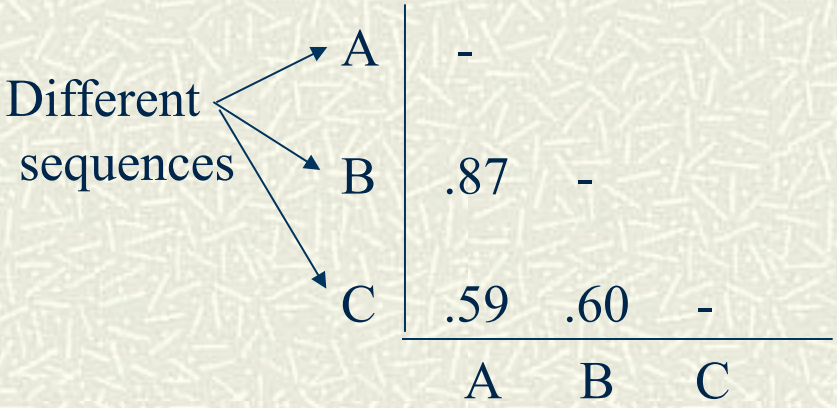

Miocene
Pliocene
Pleistocene
10
5
0
Millions of Years Before Present
Black Bear
Domestic Dog
Gray Wolf
Coyote
Cape Hunting Dog
Black-Backed Jackal
Bush Dog
Maned Wolf
Hoary Fox
Crab-Eating Fox
Gray Fox
Bat-Eared Fox
Raccoon Dog
Cape Fox
Red Fox
Fennec Fox
Kit Fox
Arctic Fox
Wolf-like canids
South American canids
Fox-like canids

Flowchart of computation steps in Clustal W (Thompson et al., 1994)



Preliminary pairwise alignments

Compare each pair of sequences.



Each number represents the number of exact matches divided by the sequence length (ignoring gaps). Thus, the higher the number the more closely related the two sequences are.

In this matrix, sequence A is 87% identical to sequence B

Step 1-Calculation of Distance Matrix

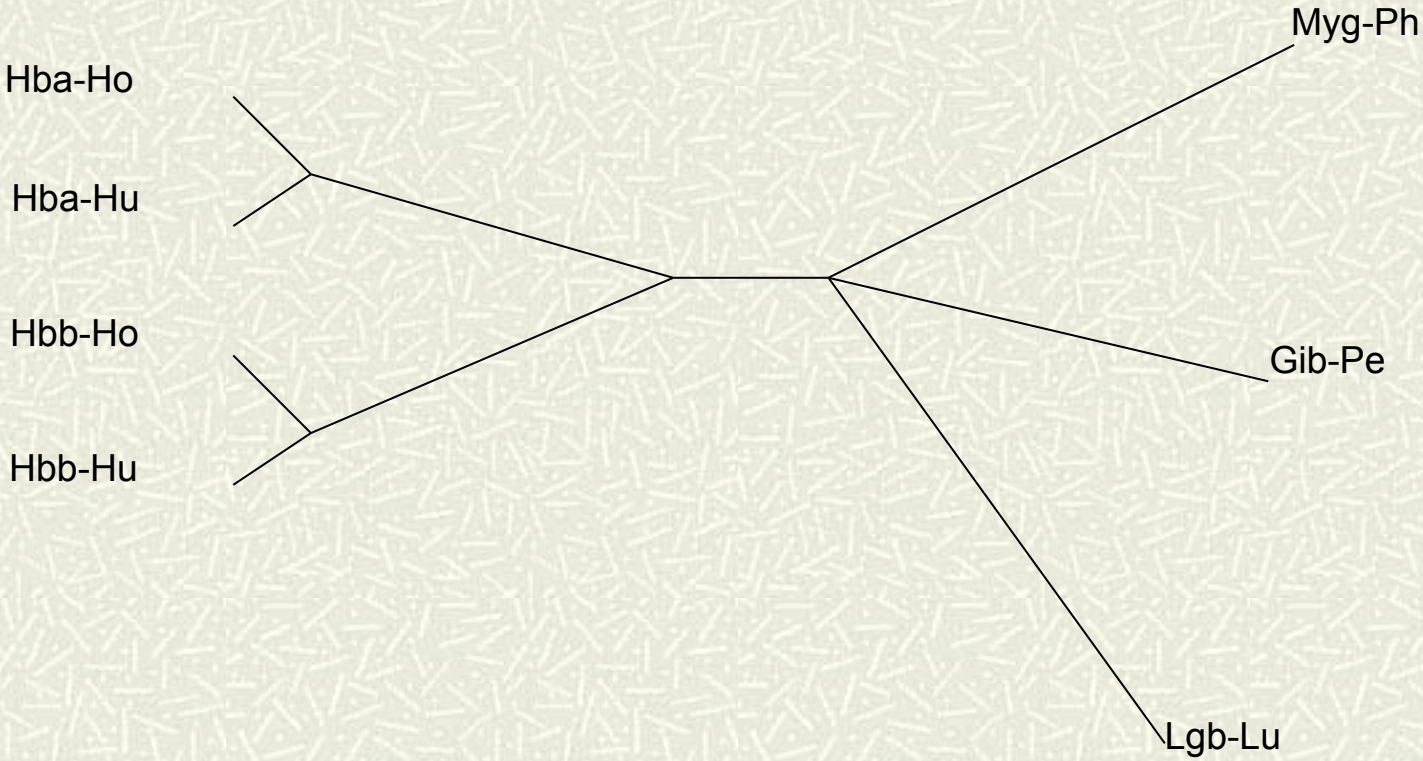
Use the Distance Matrix to create a Guide Tree to determine the "order" of the sequences.

Hbb-Hu	**1**	**-**					
Hbb-Ho	**2**	**.17**	**-**				
Hba-Hu	**3**	**.59**	**.60**	**-**			
Hba-Ho	**4**	**.59**	**.59**	**.13**	**-**		
Myg-Ph	**5**	**.77**	**.77**	**.75**	**.75**	**-**	
Gib-Pe	**6**	**.81**	**.82**	**.73**	**.74**	**.80**	**-**
Lgb-Lu	**7**	**.87**	**.86**	**.86**	**.88**	**.93**	**.90**	**-**
		1	**2**	**3**	**4**	**5**	**6**	**7**

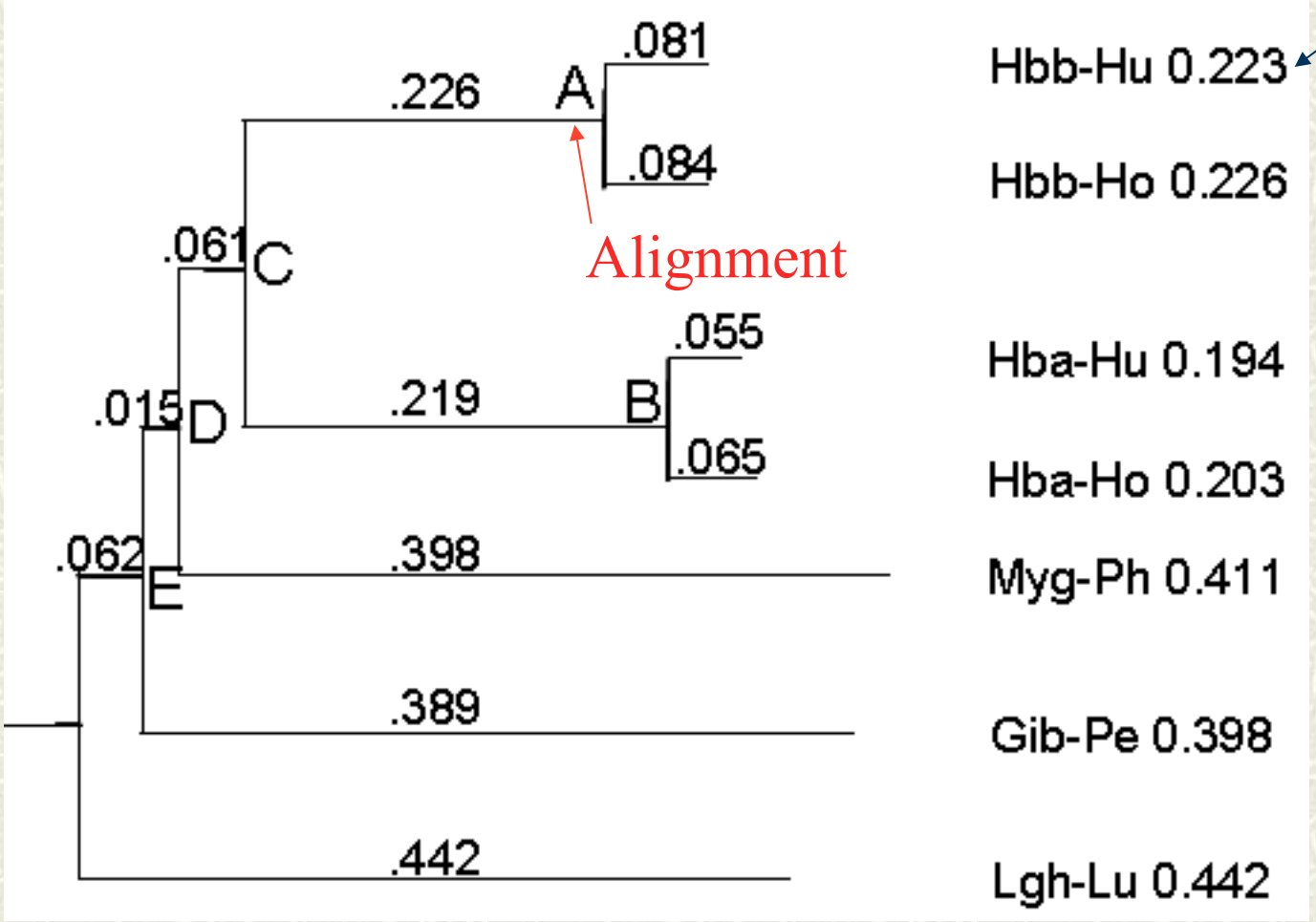
D = 1 – (I)
D = Difference score

$$I = \frac{\text{\# of identical aa's in pairwise global alignment}}{\text{total number of aa's in shortest sequence}}$$

Step 2-Create an unrooted NJ tree

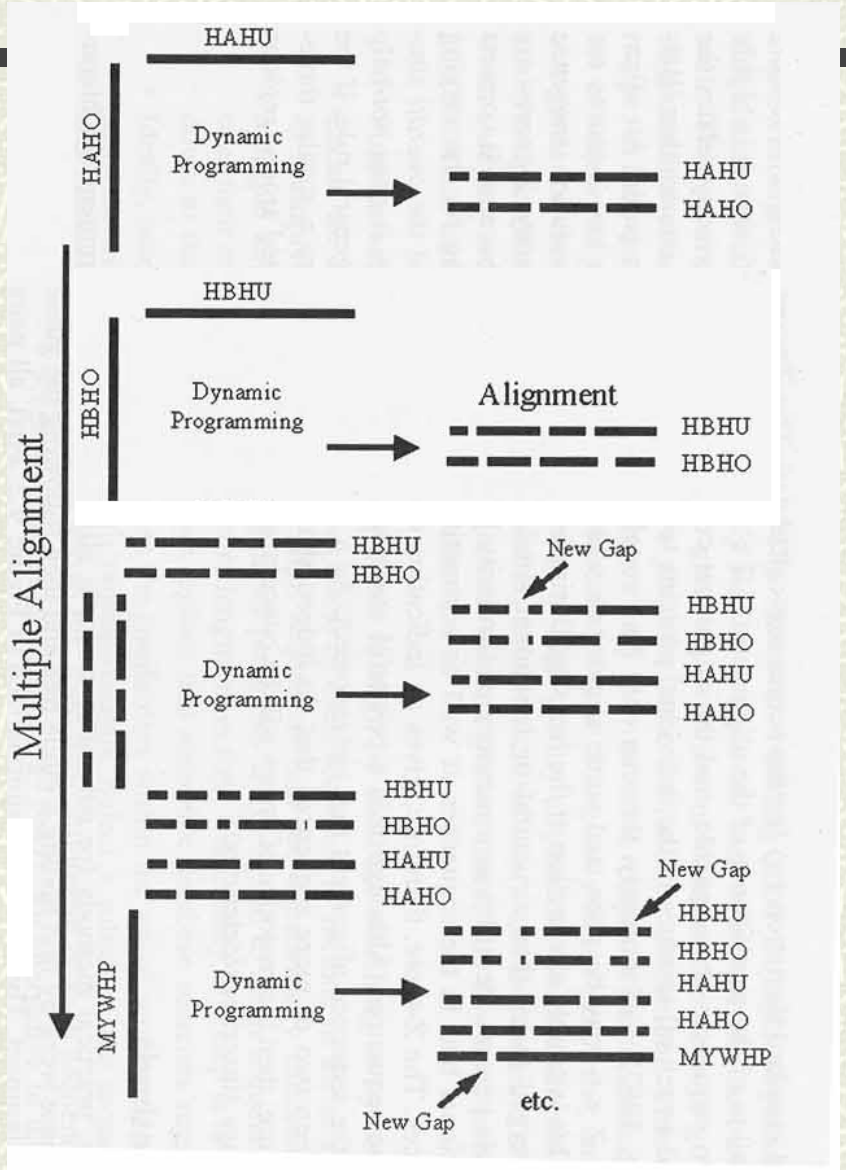


Step 3-Create Rooted NJ Tree

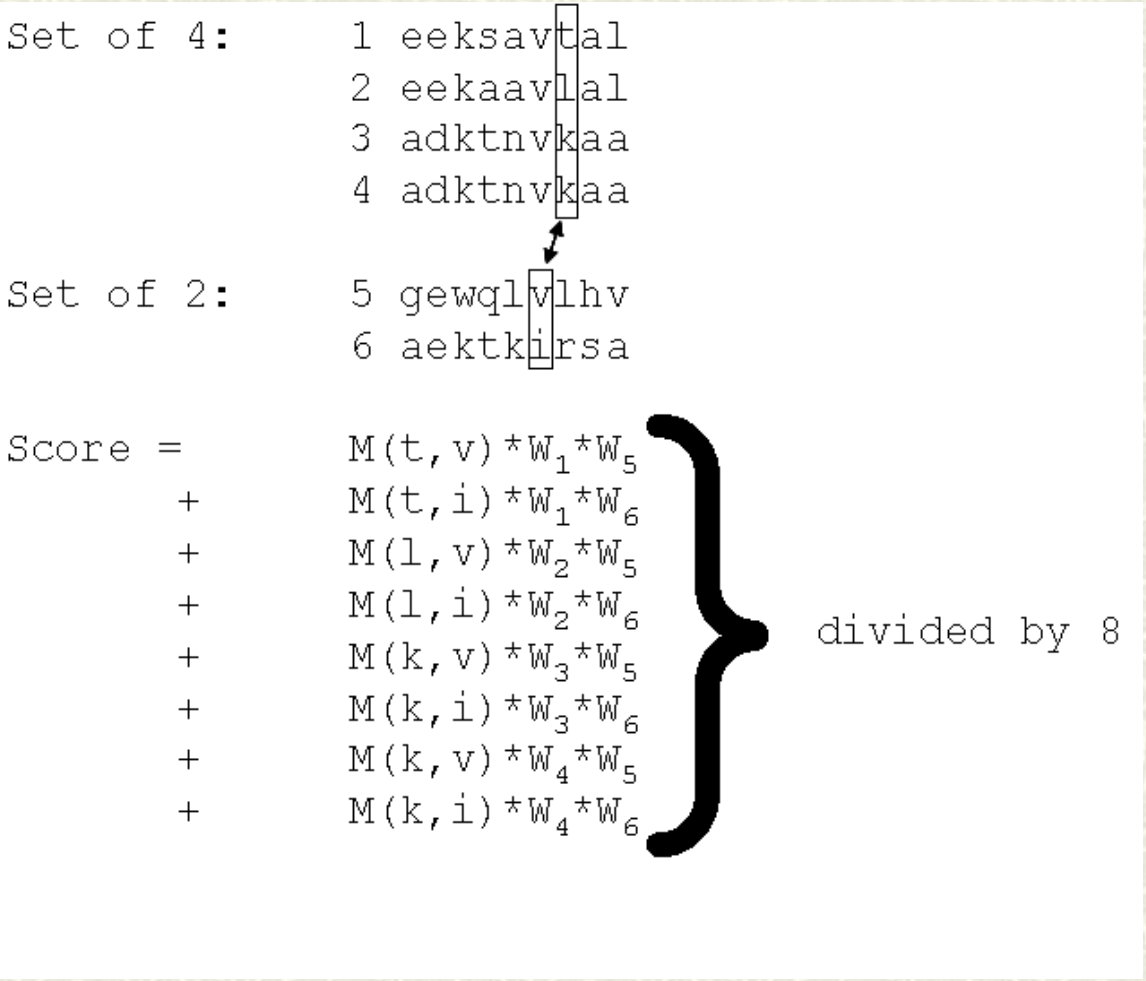


Order of alignment:
1 Hba-Hu vs Hba-Ho
2 Hbb-Hu vs Hbb-Ho
3 A vs B
4 Myg-Ph vs C
5 Gib-Pe vs D
6 Lgh-Lu vs E

Step 4-Progressive alignment



Step 4-Progressive alignment



Scoring during progressive alignment

Rules for alignment

- Short stretches of 5 hydrophilic residues often indicate loop or random coil regions (not essential for structure) and therefore gap penalties are reduced reduced for such stretches.
- Gap penalties for closely related sequences are lowered compared to more distantly related sequences ("once a gap always a gap" rule). It is thought that those gaps occur in regions that do not disrupt the structure or function.
- Alignments of proteins of known structure show that proteins gaps do not occur more frequently than every eight residues. Therefore penalties for gaps increase when required at 8 residues or less for alignment. This gives a lower alignment score in that region.
- A gap weight is assigned after each aa according the frequency that such a gap naturally occurs after that aa in nature

Amino acid weight matrices

- As we know, there are many scoring matrices that one can use depending on the relatedness of the aligned proteins.
- As the alignment proceeds to longer branches the aa scoring matrices are changed to more divergent scoring matrices. The length of the branch is used to determine which matrix to use and contributes to the alignment score.

Example of Sequence Alignment using Clustal W

```
human      ---MEEPQSDPSVEP-PLSQETFS 20
monkey     ---MEEPQSDPSIEP-PLSQETFS 20
mouse      MTAMEESQSDISLEL-PLSQETFS 23
rat        ---MEDSQSDMSIEL-PLSQETFS 20
xenopus    ---ME-PSSETGMDP-PLSQETES 19
chicken    ---MA-EEMEPLLEPTEVFMDLW- 19
                  *    . :  ::   :  :
```

Asterisk represents identity
: represents high similarity
. represents low similarity

Multiple Alignment Considerations

- **Quality of guide tree. It would be good to have a set of closely related sequences in the alignment to set the pattern for more divergent sequences.**
- **If the initial alignments have a problem, the problem is magnified in subsequent steps.**
- **CLUSTAL W is best when aligning sequences that are related to each other over their entire lengths**
- **Do not use when there are variable N- and C- terminal regions**
- **If protein is enriched for G,P,S,N,Q,E,K,R then these residues should be removed from gap penalty list. (what types of residues are these?)**

Reference: http://www-igbmc.u-strasbg.fr/BioInfo/ClustalW/