Multiple sequence alignment

Paul Magwene

How do we align multiple sequences?

- Couldn't we use a dynamic programming approach like the Needleman-Wunsch algorithm but for more than two sequences?
  - Considerations:
    - For pairwise alignment of two sequences with length $\sim L$ the time complexity of NW algorithm is $O(L^2)$
    - Turns out we can implement a dynamic programming algorithm for multiple sequences, but time complexity is $O(L^N)$ where N is the number of sequences we're aligning
    - How bad is this?
      - Do some quick estimates yourself -- Assume each operation takes 1 μs (10^{-6} seconds). How long would it take to align two 100 aa sequences using NW algorithm? How long would it take to do the multiple alignment of 10 100aa sequences using the DP approach?

Heuristic Approaches to Multiple Sequence Alignment (MSA)

- Because of the time complexity of optimal MSA, all practical algorithms for multiple sequence alignment use "heuristic" approaches (strategies that will get you a good solution, if not provable the best solution)

- A particularly popular class of heuristic approaches is based on an idea called "progressive alignment". This lies at the core of several of the most widely used MSA software tools such as ClustalW, T-COFFEE, and MUSCLE

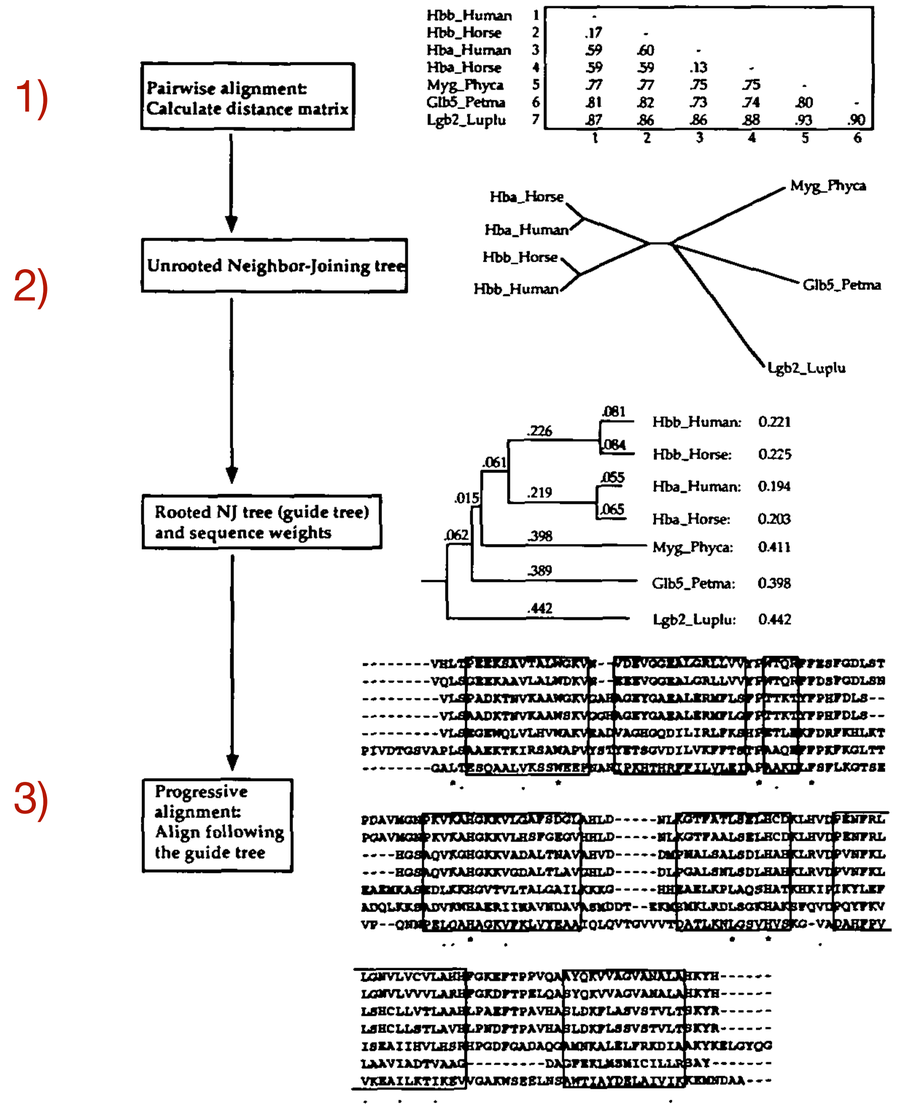
CLUSTALW algorithm for MSA

Nucleic Acids Research, 1994, Vol. 22, No. 22 4673–4680

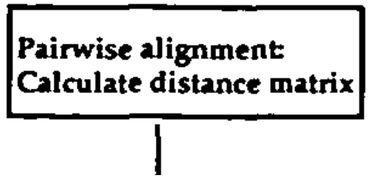
CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice

Julie D.Thompson, Desmond G.Higgins+ and Toby J.Gibson*
European Molecular Biology Laboratory, Postfach 102209, Meyerhofstrasse 1, D-69012 Heidelberg, Germany

Interesting tidbit:
This paper is one of the 10 most cited papers of all time (>40K citations as of 2014). See: https://www.nature.com/news/the-top-100-papers-1.16224



Pairwise alignments to distance matrix



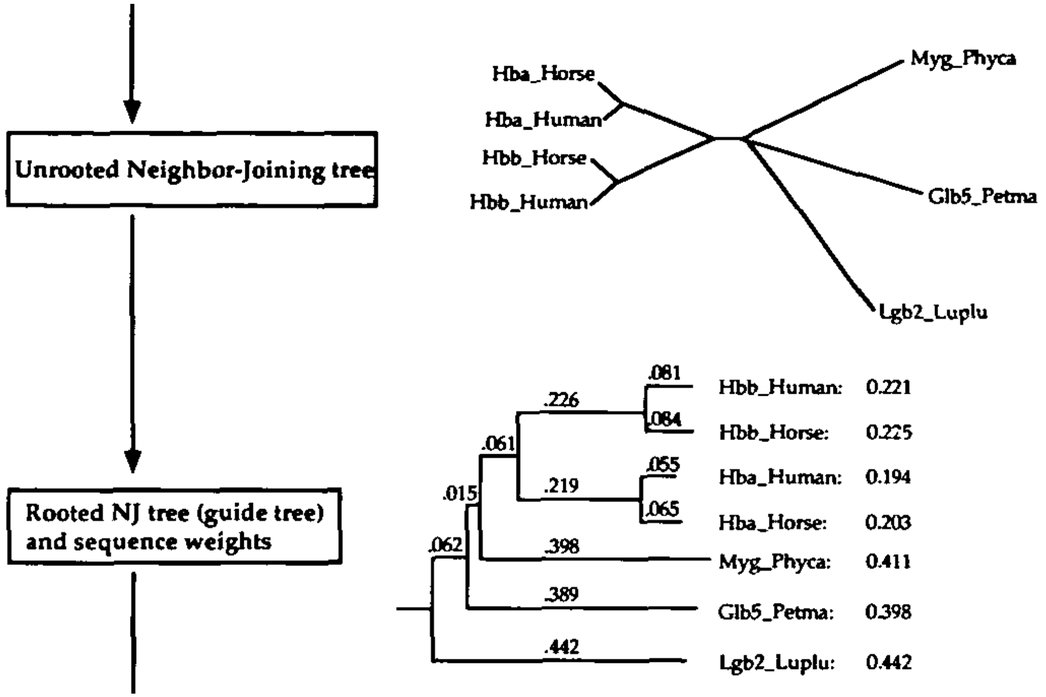
Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90
		1	2	3	4	5	6

$$d_{ij} = 1 - \frac{\%\ \text{Identity}}{100}$$

%Identity excludes gaps

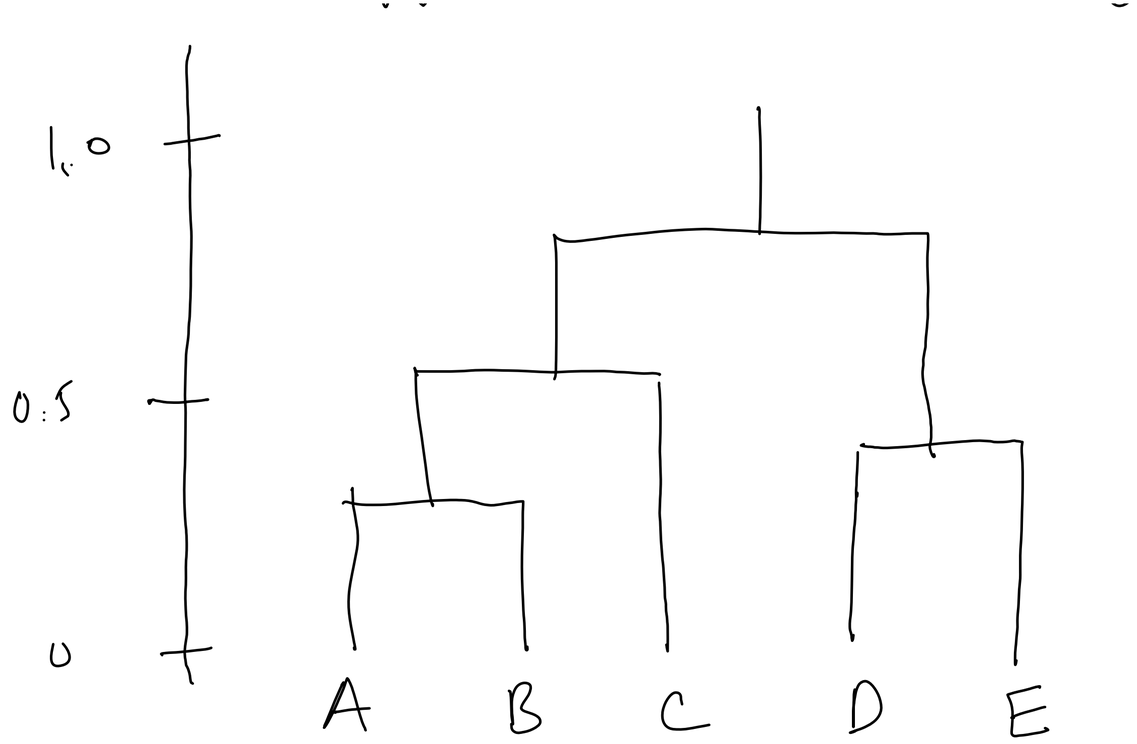
Building a guide tree using hierarchical clustering

- ClustalW uses Neighbor Joining method, but we'll illustrate with UPGMA



Hierarchical clustering

- For n object define a set of n-1 joins that represent groupings of those objects at different levels of similarity



Generic Algorithm for Agglomerative Hierarchical Clustering

1. Calculate a dissimilarity matrix for the n items
2. Join the two nearest items, i and j
3. Delete the i-th and j-th rows and columns of the dissimilarity matrix; and a new row/column that represents the dissimilarity of a new group (i,j) to all other items
4. Repeat from step 2 until there is a single group

Key Point

The different hierarchical clustering methods are determined by the function used to calculate the distance between groups in step 3.

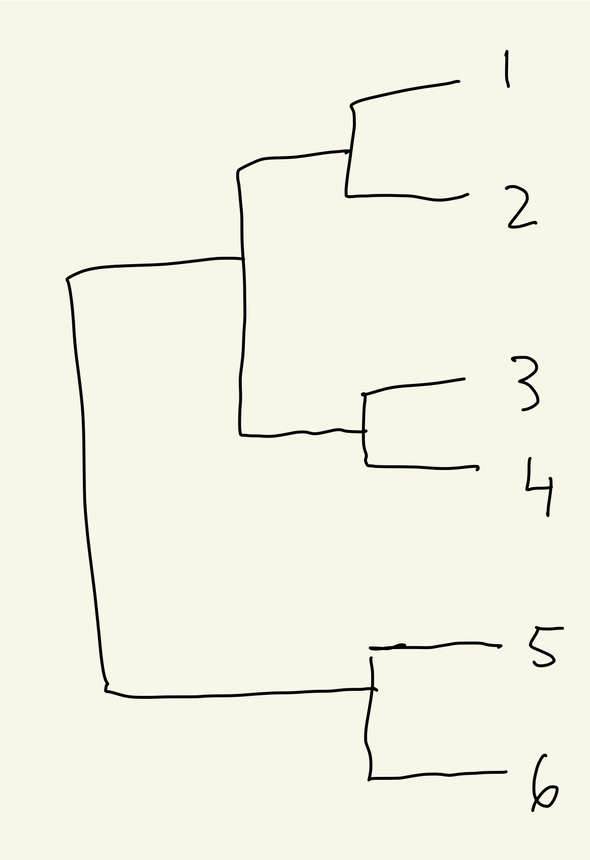
Unweighted Pair Group Method with Arithmetic mean (UPGMA)

- The distance between any two clusters A and B is taken to be the average of all distances $d_{x,y}$ between pairs of objects, x ($x \in A$) and y ($y \in B$), in A and B (mean distance between elements in each cluster)

Simple UPGMA example

	A	B	C	D	E
A	0				
B	4	0			
C	1	4	0		
D	4	2	4	0	
E	5	5	3	4	0

Carry out alignments according to the guide tree



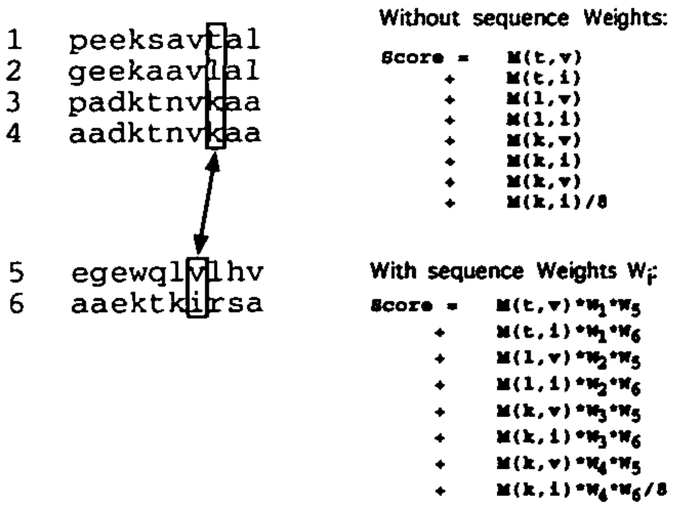


Figure 2. The scoring scheme for comparing two positions from two alignments. Two sections of alignment with 4 and 2 sequences respectively are shown. The score of the position with amino acids T,L,K,K versus the position with amino acids V and I is given with and without sequence weights. $M(X,Y)$ is the weight matrix entry for amino acid X versus amino acid Y. W_n is the weight for sequence n.

Download and install Jalview

- https://www.jalview.org/