Multiple sequence alignment

Paul Magwene

How do we align multiple sequences?

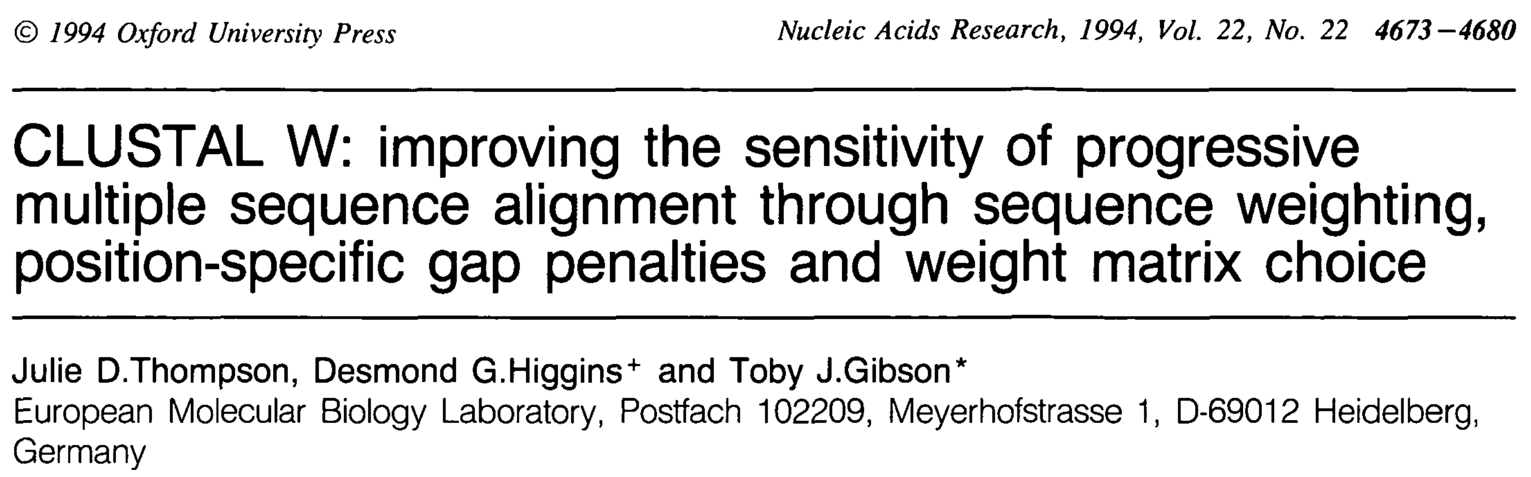
- Couldn't we use a dynamic programming approach like the Needleman-Wunsch algorithm but for more than two sequences?
    - Considerations:
        - For pairwise alignment of two sequences with length $\sim L$ the time complexity of NW algorithm is $O(L^2)$
        - Turns out we can implement a dynamic 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

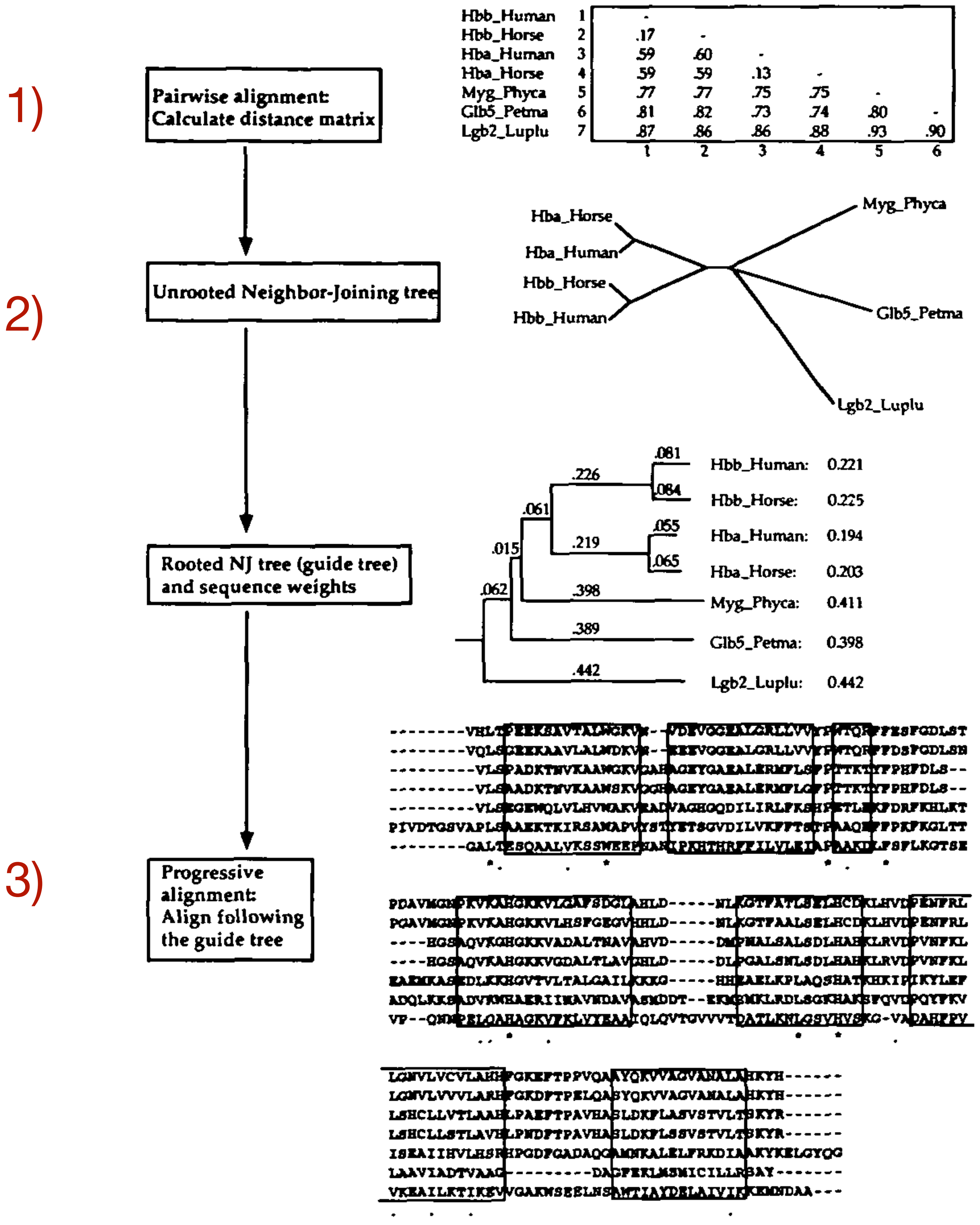


© 1994 Oxford University Press   *Nucleic Acids Research, 1994, Vol. 22, No. 22 4673–4680*

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



1) Pairwise alignment: Calculate distance matrix

2) Unrooted Neighbor-Joining tree

Rooted NJ tree (guide tree) and sequence weights

3) 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

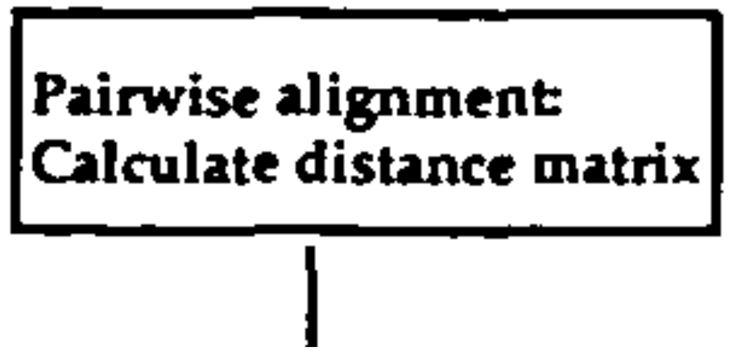
Hbb_Human:	0.221
Hbb_Horse:	0.225
Hba_Human:	0.194
Hba_Horse:	0.203
Myg_Phyca:	0.411
Glb5_Petma:	0.398
Lgb2_Luplu:	0.442

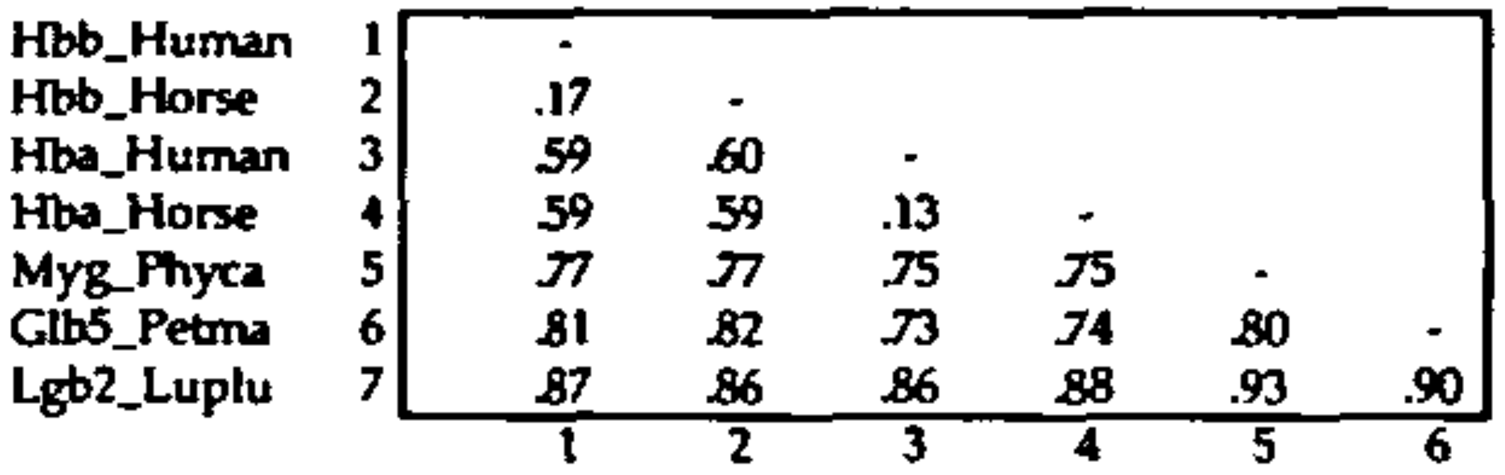
```
--------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
--------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS--
---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLS--
---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
--------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
            *.      .   *                         *.  *

PDAVMGNPKVKAHGKKVLGAFSDGLAHLD-----NLKGTFATLSELHCDKLHVDPENFRL
PGAVMGNPKVKAHGKKVLHSFGEGVHHLD-----NLKGTFAALSELHCDKLHVDPENFRL
----HGSAQVKGHGKKVADALTNAVAHVD-----DMPNALSALSDLHAHKLRVDPVNFKL
----HGSAQVKAHGKKVGDALTLAVGHLD-----DLPGALSNLSDLHAHKLRVDPVNFKL
EAEMKASEDLKKHGVTVLTALGAILKKKG-----HHEAELKPLAQSHATKHKIPIKYLEF
ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VAPAHFPV
        .. *   .                     *   *     .

LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
LAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
.  .  .                    .
```

Pairwise alignments to distance matrix





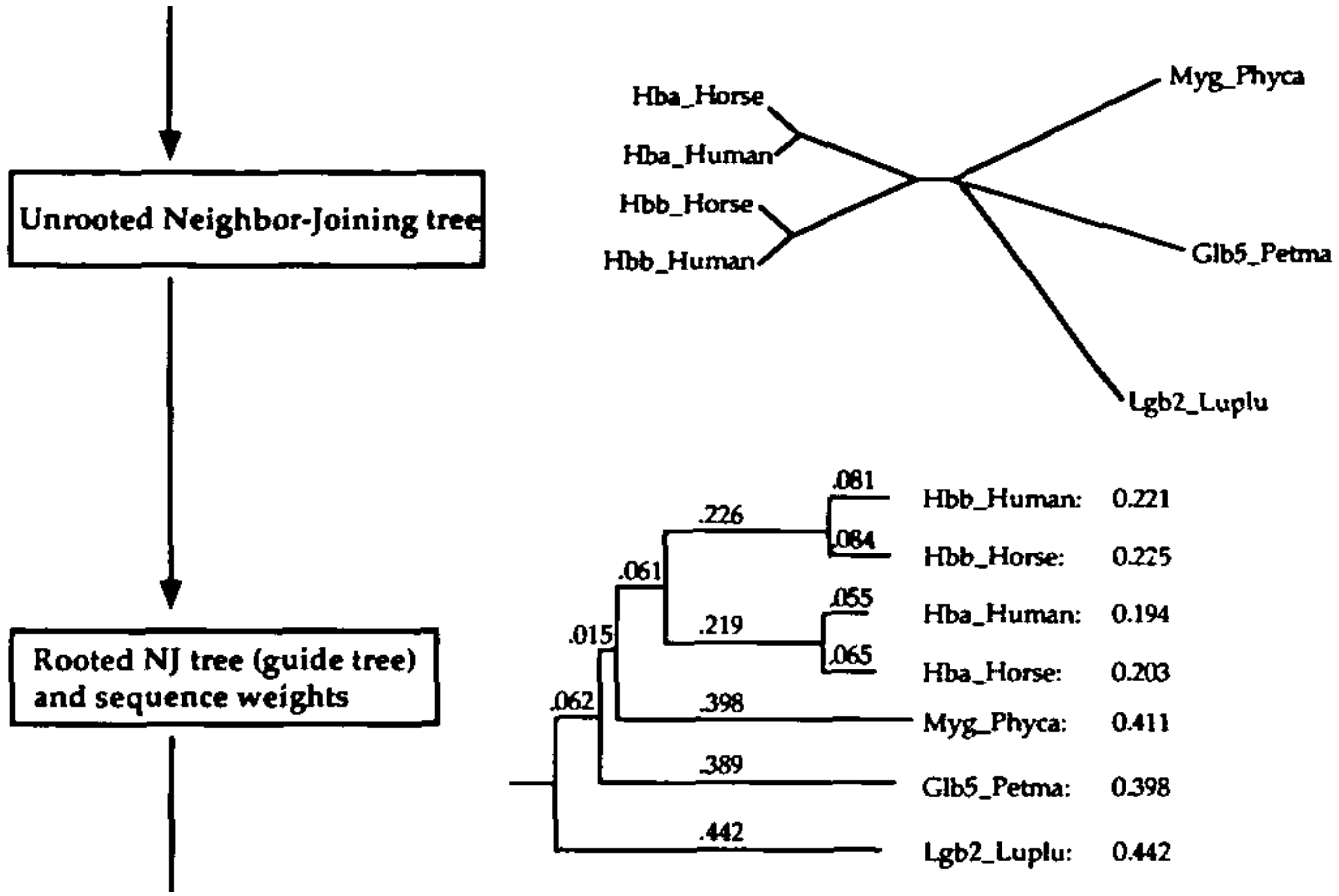
Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90
		1	2	3	4	5	6

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

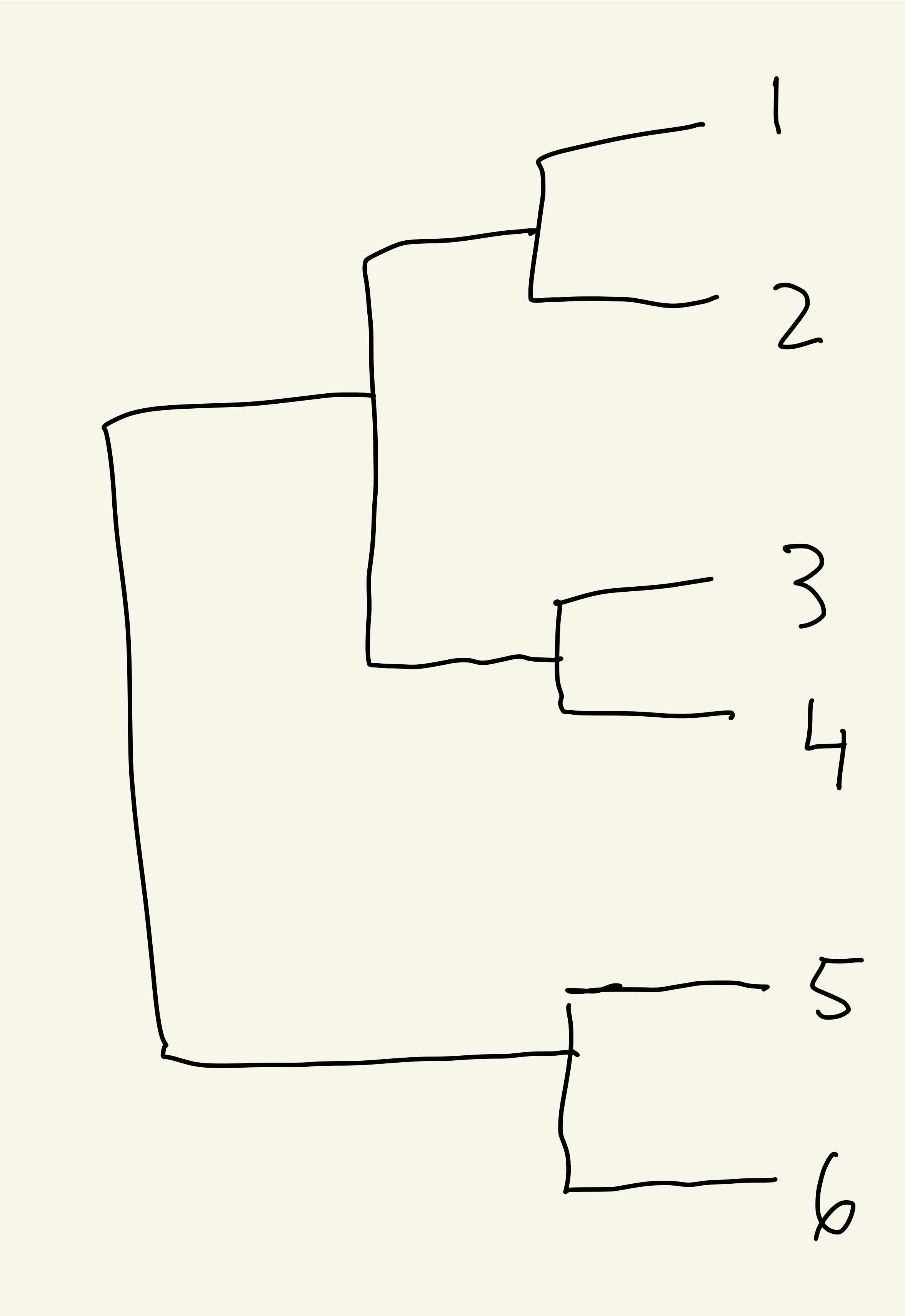
%Identity excludes gaps

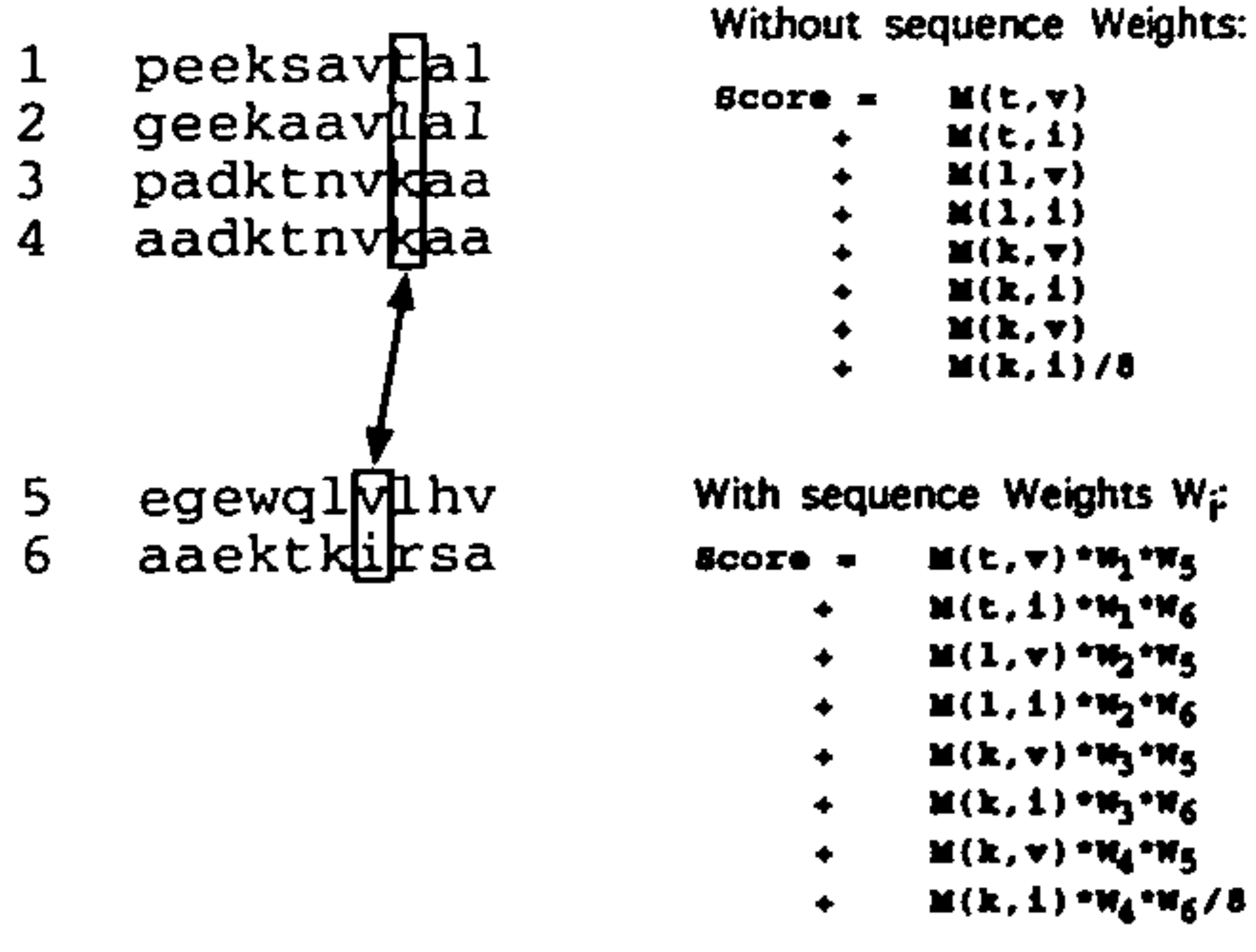
Building a guide tree using hierarchical clustering

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



Carry out alignments according to the guide tree





```
1   peeksavtal
2   geekaavlal
3   padktnvkaa
4   aadktnvkaa

5   egewqlvlhv
6   aaektkirsa
```

Without sequence Weights:

$$
\begin{aligned}
\text{Score} = \; & M(t,v) \\
+ \; & M(t,i) \\
+ \; & M(l,v) \\
+ \; & M(l,i) \\
+ \; & M(k,v) \\
+ \; & M(k,i) \\
+ \; & M(k,v) \\
+ \; & M(k,i)/8
\end{aligned}
$$

With sequence Weights W_i:

$$
\begin{aligned}
\text{Score} = \; & M(t,v)*W_1*W_5 \\
+ \; & M(t,i)*W_1*W_6 \\
+ \; & M(l,v)*W_2*W_5 \\
+ \; & M(l,i)*W_2*W_6 \\
+ \; & M(k,v)*W_3*W_5 \\
+ \; & M(k,i)*W_3*W_6 \\
+ \; & M(k,v)*W_4*W_5 \\
+ \; & M(k,i)*W_4*W_6/8
\end{aligned}
$$

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*.