CLUSTAL W Method for Multiple Alignment

Multiple Alignment: What, Why and How?

In class, we've discussed pair-wise alignment, through which we examine the similarities of two sequences by searching for the alignment with the highest score. Then we may find clue of the function of an unknown/new sequence from that of a similar but well-defined one.

Multiply alignment is an alignment with more than 2 sequences. Simple? More than that! Even we only care about the similarities of two sequences, including more sequences and performing a multiple alignment always improve the accuracy, as well as revealing more conserved patterns [1]. Moreover, for a new/unknown sequence, multiply alignments can detect homology between it and existing families of sequences, as well as predict structure and function for that sequence. Multiply alignments also provide basis for many sequence-searching algorithms such as Profile [2], PRINT [3] etc.

There are many algorithm as well as software available on line to carry out multiple alignment. Each has its advantage and limitation and applies for a particular range of sequence sets. The basic methods can be classified into two types: hierarchical methods and non-hierarchical methods. Unlike the latter, hierarchical methods don't guarantee finding the mathematically optimal alignment for an entire set of sequences, but in practice the optimal alignment found by them is good enough to make biological sense [4]. The CLUSTAL W method is a most popular, accurate and practical method in the category of hierarchical methods.

Hierarchical/progressive approaches

Since homologous sequences are evolutionarily related, we can first build a guide tree of these sequences by their pair-similarities and then follow the tree to carry out the multiple alignment of the entire set. An example of this process is shown in Fig1 [5] and the stages can be summarized as below:

1) All pairs of sequences to be aligned are compared by pair-wise alignment and a score matrix of distance or similarity is produced, indicating the divergence/similarities of each pair.
2) A guiding tree is built from the score matrix with branch length proportional to the score of each pair. In this example, NJ method is used to build the unrooted and rooted tree.
3) Multiple alignment is carried out by starting with the closest related pairs, aligning them and then including other more distant pairs progressively according to the branching order in the guide tree. Gaps present in previous alignment are fixed later on.

The progressive approach, first proposed by Feng and Doolittle [6] works efficiently and the quality of the alignments is usually excellent and reliable as long as the sequences are not too divergent. However, the choice of *alignment parameters* remains a major problem for this approach [5]. Traditionally, for a multiple alignment, one weight matrix and two gap penalties (for gap opening and extension respectively) are chosen and fixed at the alignment process. However, for very divergent sequences, different choice of these parameters may greatly affect the final solution of the alignment: (1) all residue matrices give most weight to identities, which guarantees they will find approximately the correct solution for closely related sequences. On the other hand, the scores given to mismatch will be critically important for highly divergent sequences and we have to choose appropriate matrices to get the truly optimal alignment. (2) For divergent sequences, the range of gap penalty values is very

narrow and may vary with the degree of divergence. Thus applying the same penalty at different alignment stage will cause inaccurate results. (3) Gaps occur with different probability at different positions, for example, factors such as hydrophilic stretches and residue specificities should be taken into consideration to modify the position-specific gap penalty.

CLUSTAL W Method

To solve the problem of the choice of parameters, J.D. Thompson et. al. proposed a series of modifications to the progressive alignment and the resulting program CLUSTAL W has become one of the most popular and practical tools for multiple sequence alignment. Their original paper (ref [5]) has been cited as frequently as *6768* times since its publication in1994, according to citation reports on *webofscience.com*. Below, the improvements by their methods are summarized as

(1) Individual weight for each sequence based on the guide tree;

(2) Initial gap penalties based on weight matrix, similarity and length of the sequences;

(3) Sensible local gap penalties based on existence of previous gaps and hydrophilic stretches, and residue specificities;

(4) Choice of weight matrices at different stage of alignment;

(5) Delaying the addition of very divergent sequences until the end of the alignment process.

A typical interface of CLUSTAL W is shown in Fig2 [7]. Users are allowed to choose alignment methods (accurate or fast), initial gap opening and extension penalties (0.0-100.0), weight matrices (Pam or BLOSUM) and a cut off percentage of identity (0-100%). There are also options whether to turn on Hydrophilic gaps, residue-specific gap penalties and end gap penalties.

Sequence Weighting

Individual weight is assigned to each sequence according the guide tree. Groups of closely related sequences contain similar information and thus are down weighted; on the other hand, groups of high divergent sequences receive high weights. All the weights have been normalized to be no bigger than 1, as shown in Fig 1. These weights are used as multiplication factor for scoring in a partial alignment as illustrated in Fig3. I think it quite reasonable because this modification avoids overweight of redundant information from near-duplicate sequences.

Initial Gap Penalty

The gap penalty (GP) for gap length d can be calculated from the gap opening penalty (GOP) and the gap penalty for extension:

$$GP = GOP + GEP * (d - 1)$$

Usually, users can set the initial gap values (ref Fig 2) from a range given by the program. Then the program will automatically modify these values for the alignment based on following factors.

Dependence on the weight matrix

Different weight matrices will be applied at different stages of the alignment (see the part of Weight Matrices) and it has been found that varying the gap penalties with different weight

matrices will improve the accuracy of the alignments [8]. Thus the off diagonal values of the weight matrix are added up to give the average residue mismatch score as a scaling factor for GOP.

Dependence on the similarity of the sequence

The percentage identity of two sequences or pre-aligned sequences is used as another scaling factor to disfavor the opening gaps in closely related sequences.

Dependence on the sequences lengths

As the score grows with the length of the sequences, so should the gap-opening penalty. The logarithm of the shorter length of the two sequences is used to increase the GOP with sequence length.

Dependence on the length difference

The difference between the lengths of the sequences is taken into consideration to modify the GEP so as to inhibit too many long gaps in the shorter sequence.

Thus, the initial gap penalties modified by the program are given as below (N and M are the lengths of the two sequences to be aligned):

$$GOP -> GOP + Log(Min(N,M)) * \text{(average residue mismatch score)} * \text{(percent identity)}$$

$$GEP -> GEP * (1.0 + |Log(N/M)|)$$

Position-Specific Gap Penalty

Previous programs used to apply the initial gap penalties equally at every position in the sequences, except for the terminal gaps. However, it is more reasonable to take the position specificity into consideration and assign gap penalties for every position in the two groups to be aligned. In CLUSTAL W, the modification is carried out in a hierarchical manner.

(1) If there is already a gap at a position from earlier alignment, the GOP and GEP are reduced to favor gaps there: $GOP -> GOP * 0.3 * (\text{no. of sequences without a gap/no. of sequences})$

(2) If condition (1) is not satisfied but the position is within 8 residues of an existing gap, the GOP is increased to discourage gaps to get too close: $GOP -> GOP * (2 + \text{(8-distance from gap)*2)/8})$

(3) Short stretches of hydrophilic residues usually indicate loop or random coil region and encourage gaps more than the regular secondary structure. Thus, if conditions (1) and (2) are not satisfied, GOP is reduced by one third at such positions.

(4) If none of the above conditions are met, the GOP is adjusted by the residue specificity (Table1), derived from the original table of relative frequencies of gaps adjacent to each residue [9].

An example of the application of the above rules is shown in Fig 4.

Weight Matrices

The mostly used weight matrices are the PAM and BLOSUM series. Users have the option to cho0se from these two series. For each series, there is a range of matrices appropriate for sequences of different degree divergence. Depending on the distances measured from the guide tree, different types of matrices are used, for detail, see ref [5].

Divergent sequences

The more divergent the sequence is, the more difficult it is to align it correctly. If we delay the including of the most divergent sequences until the end of the alignment process, we may stand a better chance to correctly place and gaps and match the weakly conserver region with the rest. Users can choose the cut-off identity percentage to the delay the addition of divergent sequences. The default value of CLUSTAL is 40%.

By the above modifications of parameter choice, CLUSTAL W has greatly improved the sensitivity of progressive alignment, especially for difficult alignment with very divergent sequences. However, it has its disadvantages. One limitation, common to all hierarchical methods, is that it can't guarantee a mathematical optimal solution and there is no way of quantifying whether the alignments are good or not. Another problem, which is suffered by most progressive approaches, is the local minimum problem [5]. As more sequences are progressively added to the alignment, the mistakes made in earlier alignment can't be corrected and thus carried on to the final alignment. Iterative or stochastic sampling procedures have been suggested to correct this local minimum problem.

Great efforts have been used into tuning the parameters of CLUSTAL W to produce alignments consistent with those done manually or with 3D structure comparison [4], and make the parameters appropriate for a wide range of alignment problems. For example, J.D. Thompson et.al. tested their methods by studying the case of over 60 SH3 domains with highly divergent sequences. With their default parameters and repeating the alignment process twice, they achieved an alignment almost as exactly correct as suggested by known structural information [5]. Overall, CLISTAL W is a reliable, practical and efficient tool for multiple sequences alignment, though caution should be taken to judge the significance and validity of the results.

Reference:

[1] R.B.Russell and G.J.Barton, *Proteins* **14**,309-323 (1992)

[2] M. Gribskov, A. D. McLachlan and D. Eisenberg, *Proc. Nat. Acad. Sci.*, **84**, 4355-4358(1987)

[3] T.K. Attwood et.al. *Nucl. Acids Res.* **28**, 225-227(2000)

[4] Andreas D. Baxevanis & B.F. Francis Ouellette, *Bioinformatics: a practical guide to the analysis of genes and proteins* (2001)

[5] J.D.Thompson, D.G. Higgins and T. J.Gibson, *Nucl. Acids Res.* **22,** 4673-4680 (1994)

[6] D.F.Feng and R.F. Doolittle, *J.Mol.Evol.* **25,** 351-360 (1987)

[7] http://dot.imgen.bcm.tmc.edu: 9331/multi-align/Options/clustalw.html

[8] R. Luthy, I.Xenarios and P.Bucher, *Protein Sci.***3**, 139-146 (1994)

[9] S. Pascarella and P. Argos, *J. Mol.Biol*, **224**, 461-471 (1992)

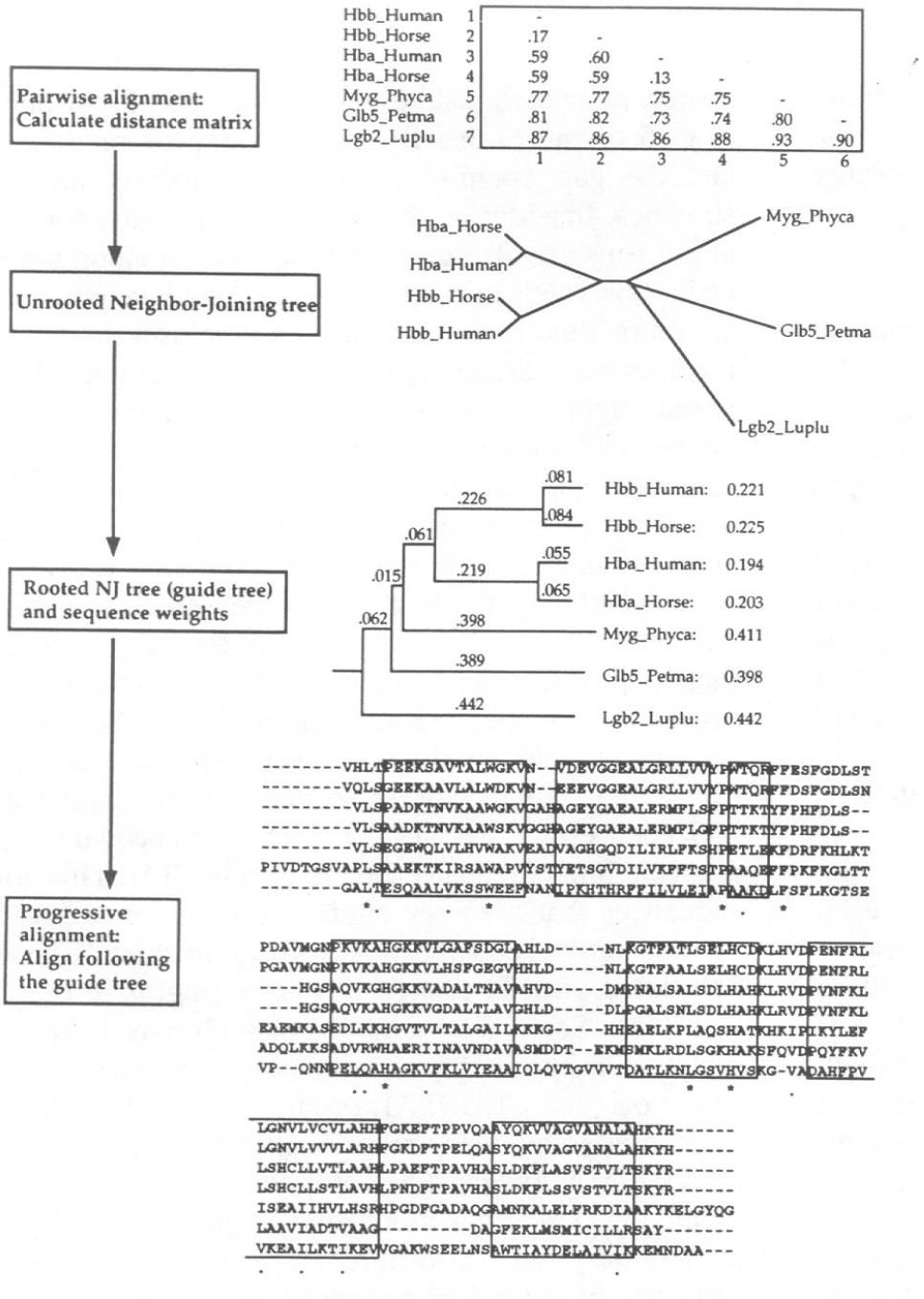

Fig 1 Basic procedure of progressive alignment[5]. A set of 7 globins of known tertiary structure is used. First, a distance matrix is given with the mean number of difference per residue of each sequence pair. Second, an unrooted tree is built with branch length proportional to the distance. A guide tree is then produced with weights calculated for each sequence (see the part of sequence weighting in text). In the multiple alignment, the approximate positions of 7 α helices common to all 7 proteins are shown.

Protein Sequence Parameters:

Fast Pair-wise Alignment Parameters:

K-tuple (word) size: 1

Window size: 5

Scoring method: Percentage

Number of top diagonals: 5

Gap penalty: 3

Multiple Alignment Parameters:

Weight matrix: blosum

Gap opening penalty: 10.0

Gap extension penalty: 0.05

Hydrophilic gaps: On

Hydrophilic residues: GPSNDQERK

Residue-specific gap penalties: On

Other Parameters:

Quicktree (always On): On

Divergence cutoff (% identity for delay): 40

Gap separation distance: 8

End gap separation penalty: Off

Output order: aligned

Fig2 A typical parameters input interface of CLUSTAL W multiple alignment [7]

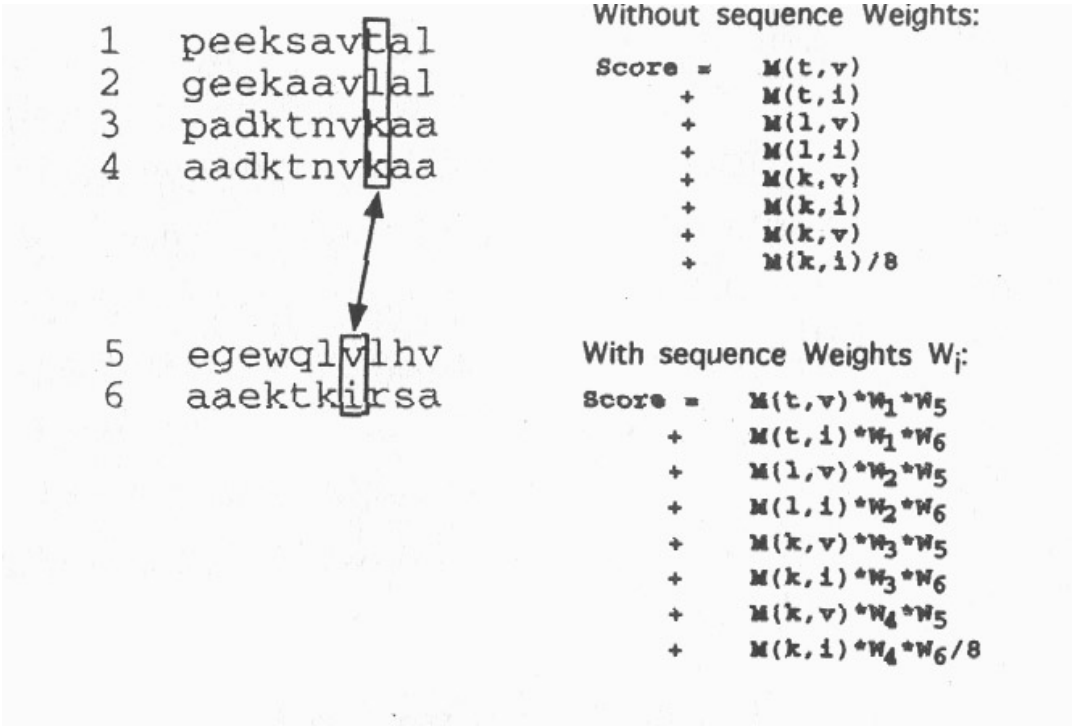

Fig 3. Scoring schemes without and with sequence weights [5]. Two pre-aligned group of sequences (1-4 and 5-6) are compared and scored for the position with amino acid T, L, K, K versus the position with V, I. M(X,Y) is the weight matrix entry for amino acid X versus Y and W_n is the weight for sequence n.

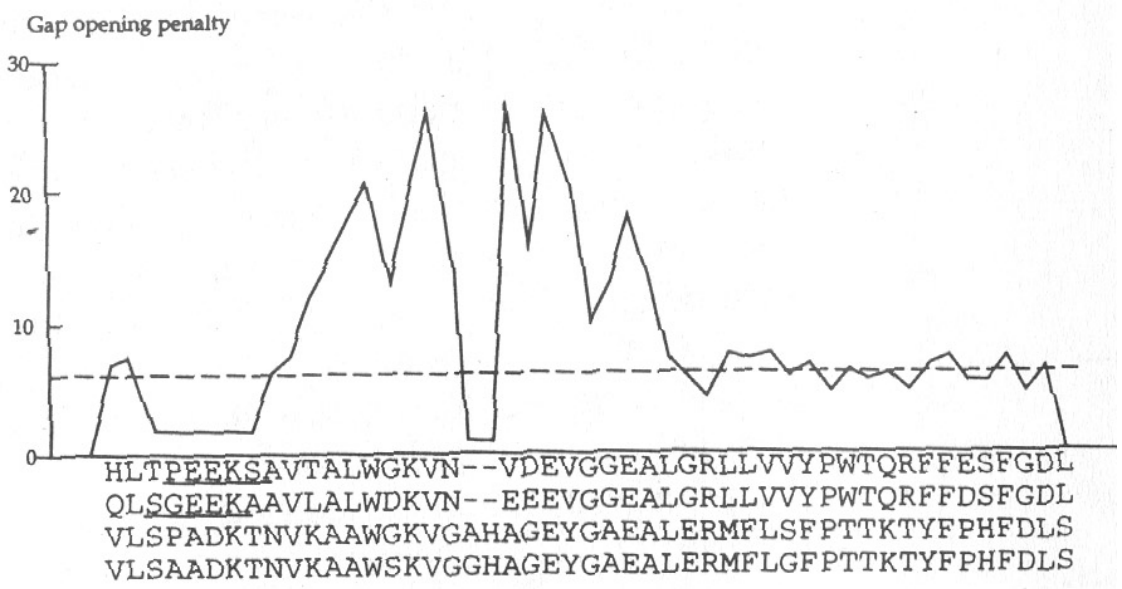

Fig 4 The position-specific GOP for a section of partial alignment [5]. The dotted line indicates the initial GOP and the unlined regions are hydrophilic stretches.

From this plot, we can see that the lowest values correspond to the end of the alignment, position with gaps in earlier alignment and the hydrophilic stretches, while the highest values correspond to gaps with 8 residues. The rest of variation is due to the adjustment by the residue specific gap penalties.

Table 1. Pascarella and Argos residue specific gap modification factors

A	1.13	M	1.29
C	1.13	N	0.63
D	0.96	P	0.74
E	1.31	Q	1.07
F	1.20	R	0.72
G	0.61	S	0.76
H	1.00	T	0.89
I	1.32	V	1.25
K	0.96	Y	1.00
L	1.21	W	1.23

Table 1 The residue-specific gap penalty modification factor [5]. The values are normalized around a mean value for residue H. The lower the value, the greater the chance for the residue to have an adjacent gap.