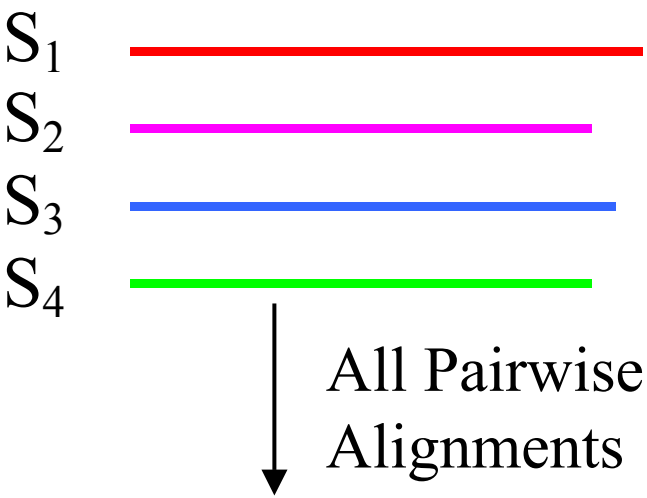
Multiple alignment: heuristics

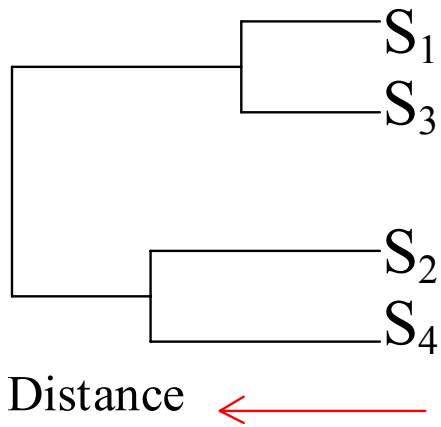
ClustalW steps



Similarity Matrix

	S_1	S_2	S_3	S_4
S_1		4	9	4
S_2			4	7
S_3				4
S_4				

Cluster Analysis



Dendrogram

Multiple Alignment Step:
1. Aligning S_1 and S_3
2. Aligning S_2 and S_4
3. Aligning (S_1,S_3) with (S_2,S_4).

From Higgins(1991) and Thompson(1994).

Example (from the text book)

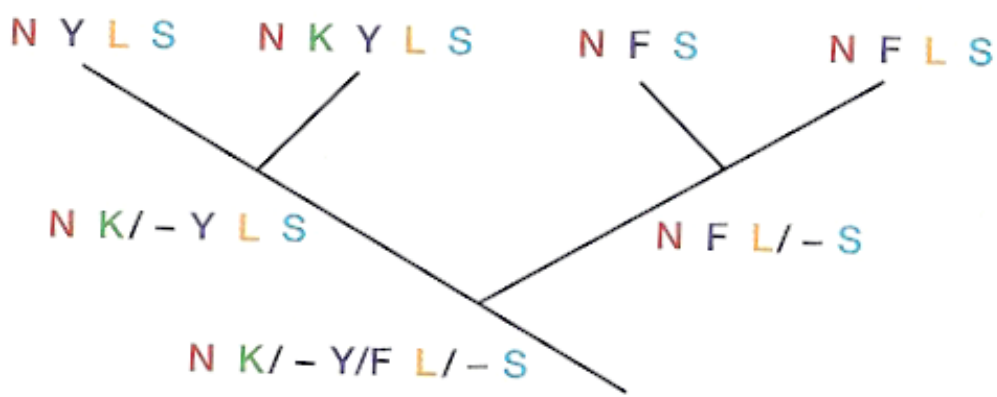


Figure 4.7. Progressive sequence alignment. Sequences are represented as the outermost branches (leaves) on an evolutionary tree. The most closely related sequences are first aligned by dynamic programming, providing a representation of ancestor sequences in deeper branches with uncertainties where amino acids have been substituted or positioned opposite a gap. These sequences are the same as those shown in EVMSA. The challenge to the msa method is to utilize an appropriate combination of sequence weighting, scoring matrix, and gap penalties so that the correct series of evolutionary changes may be found.

Sum of Pairs (SP) method

- Consider aligning the following 4 protein sequences

  S1 = AQPILLLV

  S2 = ALRLL

  S3 = AKILLL

  S4 = CPPVLILV

- Next consider the following MSA

```
A Q P I L L L V
A L R - L L - -
A K - I L L L -
C P P V L I L V
```

Sum of Pairs cont.

- Assume c(match) = 1 , c(mismatch) = -1 , and c(gap) = -2 , also assume c(-, -) = 0 to prevent the double counting of gaps.
- Then the SP score for the 4th column of the MSA would be $SP(m_4) = SP(I, -, I, V)$
  $= c(I,-) + c(I,I) + c(I,V) + c(-,I) + c(-, V) + c(I,V)$
  $= -2 + 1 + (-1) + (-2) + (-2) + (-1)$
  $= -7$
- To find SP(MSA) we would find the score of each m_i and then SUM all $SP(m_i)$ scores to get the score MSA.
- To find the optimal score using this method we need to consider all possible MSA.

Some Problems with the SP Method

- Consider column 1 of our example ie A,A,A,C for this column we get $SP(m_4) = SP(A,A,A,C)$
$= 1 + 1 + (-1) + 1 + (-1) + (-1)$
$= 0$
whereas if we had A,A,A,A we get a score of $SP(A,A,A,A) = 1+1+1+1+1+1 = 6$, thus we get a difference of 6 for what could be explained by a single mutation.
- The SP method tends to overweight the influence of mutations
- By introducing differential weights to individual sequences, this problem can be circumvented.

The ClustalW method

- ClustalW is a progressive method for MSA
- The term “progressive” is used because ClustalW starts with using a pairwise method to determine the most related sequences and then progressively adding less related sequences or groups of sequences to the initial alignment.
- Clustal comes in 3 versions

  Clustal - gives equal weight to all sequences

  ClustalW - has the ability to give different weights to the sequences

  ClustalX - provides a graphical interface to ClustalW

The ClustalW algorithm

- Step 1 : Determine all pairwise alignment between sequences and determine degrees of similarity between each pair.
- Step 2 : Construct a similarity tree.
- Step 3 : Combine the alignments starting from the most closely related groups to the most distantly related groups, using the “once a gap always a gap” rule .

ClustalW Step 1

- Use a pairwise alignment method to compute pairwise alignment amongst the sequences.
- Using the pairwise alignments compute a “distance” between all pairs of sequences. A method commonly used is as follows:

  for each pairwise alignment look at the non-gapped position and count the number of mismatches between the two sequences, then divide this value by the number of non-gapped pairs to calculate the distance, for example

```
NKL-ON      distance = 1/4 = .25
-MLNON
```

ClustalW Step 1 continued

- After computing the "distance" between all pairs of sequences we put them in to a matrix. For example if we consider a set of 7 sequences we could have the following matrix:

Seq.	S1	S2	S3	S4	S5	S6	S7
S1	-						
S2	.17	-					
S3	.59	.60	-		(SYMMETRIC)		
S4	.59	.59	.13	-			
S5	.77	.77	.75	.75	-		
S6	.81	.82	.73	.74	.80	-	
S7	.87	.86	.86	.88	.93	.90	-

ClustalW Step 2

- Construct a similarity tree. ClustalW uses a distance matrix and the Neighbor Joining method to construct the similarity tree. The root is placed a the midpoint of the longest chain of consecutive edges.

ClustalW Step 3

- Combine the alignments starting from the most closely related groups to the most distantly related groups by going from tip of tree to the root of the tree.
- In our example we first align S1 with S2 (profile 1) then S3 with S4 (profile 2), then align grp1 with grp2, we continue until the root is reached.

  Each alignment (sequence-sequence, sequence-profile, profile-profile) involves dynamic programming by the SP score method.

ClustalW: Avoiding SP problems

In order compensate for biased representation in large subfamilies, ClustalW reduces the weight of related sequences.

To assign a weight to a sequence, ClustalW does the following:

(1) It uses the values on the NJ-tree from the sequence to the root of the tree.

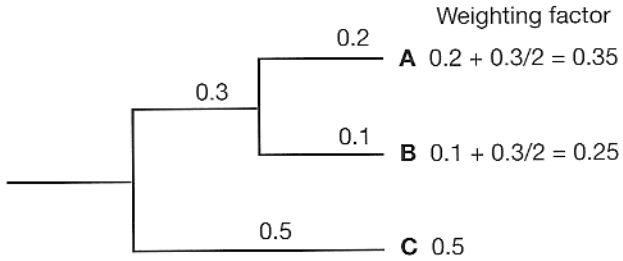
(2) If two or more sequences share a branch, which may indicate an evolutionary relationship, its value is split amongst the sequences.

Assigning weights to the sequences in scoring multiple alignment

- Each sequence has assigned a weight and the contribution of this sequence in the global alignment is scaled using this weight.
- We assume that the evolutionary tree computed in the second stage of the algorithm has an associated branch length. Then the weight of associated with the sequence is proportional to the sum of the branch length from the root to the corresponding leaf.

Example

A. Calculation of sequence weights



B. Use of sequence weights

Column in alignment 1

Sequence A (weight a) .....K.....

Sequence B (weight b) .....I.....

Column in alignment 2

Sequence C (weight c) .....L.....

Sequence D (weight d) .....V.....

Score for matching these two column in an msa =

[a x c x score (K,L) +
a x d x score (K,V) +
b x c x score (I,L) +
b x d x score (I,V)] / 4

Figure 4.8. Weighting scheme used by CLUSTALW (Higgins et al. 1996). (*A*) Sequences that arise from a unique branch deep in the tree receive a weighting factor equal to the distance from the root. Other sequences that arise from branches shared with other sequences receive a weighting factor that is less than the sum of the branch lengths from the root. For example, the length of a branch common to two sequences will only contribute one-half of that length to each sequence. Once the specific weighting factors for each sequence have been calculated, they are normalized so that the largest weight is 1. As CLUSTALW aligns sequences or groups of sequences, these fractional weights are used as multiplication factors in the calculation of alignment scores. (*B*) Illustration of using sequence weights for aligning two columns in two separate alignments. Note that this sequence weighting scheme is the opposite to that used by MSA, because the more distant a sequence from the others, the higher the weight given. For a comparison of additional weighting schemes, see Vingron and Sibbald (1993).

Protein gap parameters in ClustalW

- RESIDUE SPECIFIC PENALTIES are amino acid specific gap penalties that reduce or increase the gap opening penalties at each position in the alignment or sequence. As an example, positions that are rich in glycine are more likely to have an adjacent gap than positions that are rich in valine.
- HYDROPHILIC GAP PENALTIES are used to increase the chances of a gap within a run (5 or more residues) of hydrophilic amino acids; these are likely to be loop or random coil regions where gaps are more common.
- GAP SEPARATION DISTANCE tries to decrease the chances of gaps being too close to each other. Gaps that are less than this distance apart are penalised more than other gaps. This does not prevent close gaps; it makes them less frequent, promoting a block-like appearance of the alignment.
- END GAP SEPARATION This is useful when you wish to align fragments where end gaps are not biologically meaningful.
- The use of secondary structure-based penalties has been shown to improve the accuracy of multiple alignment. Therefore CLUSTAL W now allows gap penalty masks to be supplied with the input sequences. The masks work by raising gap penalties in specified regions (typically secondary structure elements) so that gaps are preferentially opened in the less well conserved regions (typically surface loops). Large experimentally determined lookup table for gap preferences are available.

Problems with ClustalW and other "progressive alignments"

- Dependence of the initial pairwise sequence alignment.
- Propagating errors form initial alignments.

Using MSA programs online

- http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html

Conclusion

- Heuristics to find good MSA employed by ClustalW
  - Progressive sequence alignment
  - Weighting scheme to reduce influence of similar sequences
  - Gap penalty tricks

- Problems
  - Dependence of the initial pairwise sequence alignment.
  - Propagating errors form initial alignments.

MSA version 2.0

- Program info:
  - Authors of the program and consecutive improvements:Carrillo, Lipman, Altschul, Shaffer, Gupta, Kececioglu,…
  - Good description can be found at: http://www.psc.edu/general/software/packages/msa/manual/manual.html
  - http://www.psc.edu/general/software/packages/msa/manual/manual.html