Multiple sequence alignment via dynamic programming

How to score an alignment of more than two sequences? One alternative: Sum-of-pairs criterion.

The idea is that each alignment between n sequences actually specifies $n(n-1)/2$ pairwise alignments. This 3-sequence alignment ...

```
Sequence 1  A  T  C  G  A  G  T  A
Sequence 2  A  -  C  G  T  -  T  A
Sequence 3  -  T  C  G  A  -  T  A
```

specifies 3 pairwise alignments

```
Sequence 1  A  T  C  G  A  G  T  A
Sequence 2  A  -  C  G  T  -  T  A
```

```
Sequence 1  A  T  C  G  A  G  T  A
Sequence 3  -  T  C  G  A  -  T  A
```

```
Sequence 2  A  -  C  G  T  -  T  A
Sequence 3  -  T  C  G  A  -  T  A
```

One way to score a multiple sequence alignment is to use the sum of the scores of the pairwise alignments that it specifies.

Usually, pairing two gap "-" symbols is treated as not affecting the pairwise alignment score under the sum-of-pairs criterion.

The dynamic programming algorithm is straightforward to extend to more than 2 sequences.

For 2 sequences of respective length L_1 and L_2, the dynamic programming algorithm involves traversing an $L_1 + 1$ by $L_2 + 1$ grid where scores of subalignments are stored at the nodes of the grid.

For 3 sequences, the grid would be 3 dimensional.

Instead of a L_1+1 by L_2+1 grid, we would have a L_1+1 by $L_2 + 1$ by $L_3 + 1$ grid.

Number of points in a grid for an N sequence alignment would be

$$\prod_{i=1}^{N} (L_i + 1)$$

An entry at a grid point that represents the score of an optimal N-sequence subalignment would be determined by entries at up to $2^N - 1$ neighboring grid points.

The Carillo-Lipman algorithm

Let α be an optimal similarity-based alignment between sequences $1, \ldots, N$ according to the sum-of-pairs criterion.

Let H be a (hopefully good) guess as to the optimal alignment between sequences $1, \ldots, N$.

α_{ij} (H_{ij}) will be the pairwise alignment (i.e., the projection) between sequences i and j ($i < j$) that is specified by the $N-$sequence alignment α (H).

$$\begin{aligned}
\text{score}(\alpha) &\geq \text{score}(H) \\
\sum_{i<j} \text{score}(\alpha_{ij}) &\geq \text{score}(H) \\
\text{score}(\alpha_{xy}) + \sum_{i<j,(x,y)\neq(i,j)} \text{score}(\alpha_{ij}) &\geq \text{score}(H) \\
\sum_{i<j,(x,y)\neq(i,j)} \text{score}(\alpha_{ij}) &\geq \text{score}(H) - \text{score}(\alpha_{xy})
\end{aligned}$$

Let P_{ij} be the optimal pairwise alignment between sequences i and j.

Because $\text{score}(P_{ij}) \geq \text{score}(\alpha_{ij})$ for all (i, j),

$$\sum_{i<j,(x,y)\neq(i,j)} \text{score}(P_{ij}) \geq \text{score}(H) - \text{score}(\alpha_{xy})$$

So,

$$\text{score}(\alpha_{xy}) \geq \text{score}(H) - \sum_{i<j,(x,y)\neq(i,j)} \text{score}(P_{ij})$$

A simple dynamic programming algorithm exists that can determine the score of the best scoring alignment that passes through each grid point in a pairwise alignment grid.

For a point on an N-dimensional grid, we can consider all $N(N-1)/2$ pairwise alignment projections.

If, for at least one of the grid points on these pairwise projections, the score of the best alignment between a sequence x and y that passes through the grid point is less than ...

$$\text{score}(H) - \sum_{i<j,(x,y)\neq(i,j)} \text{score}(P_{ij})$$

then we know that the optimal alignment α cannot pass through this point on the N-dimensional grid.

A careful implementation is need to make this algorithm useful...

The algorithm is implemented in the MSA program (Lipman et al. 1989) and allows optimal alignment of up to roughly 10 protein sequences.

Why should the sum-of-pairs criterion be used? Consider 2 hypothetical sites in an alignment between 5 sequences. Each site has exactly 1 isoleucine to valine change.

Sankoff and collaborators (1973, 1975, 1976) realized long ago that alignment and phylogenies are closely related.

Hogeweg and Hesper, J Mol Evol (1984) 20:175-186

General Outline of Method:

1. A Pairwise Distance-Similarity-Score is computed for each pair of sequences

2. The pairwise score matrix is used by some distance matrix method (choose your own) to construct a tree.

3. "The sequences are successively pairwise-aligned, following the branches of the tree, and internode sequences are constructed." If the alignment position contains a mismatch or gap, the decision about the content of the internode sequence is delayed until next sequence is added to tree.

4. From the initial multiple sequence alignment, again compute pairwise distance-similarity-score for each sequence pair. If some alignment positions are much more variable than others, differential weighting of alignment positions can be considered.

5. Goto Step 3, continue until process converges.

Feng and Doolittle, J Mol Evol (1987) 25:351-360.

Used as basis of CLUSTALW (Thompson et al. 1994) program.

"Once a gap, always a gap" : Closely related sequence pairs should be used to position a gap because they have more reliable information.

1. Distance Matrices are constructed from pairwise alignment scores. A distance method is used to infer tree shape.

2. Sequences are progressively aligned according to branches of the distance tree. When a gap or mismatch occurs in more than 25 percent of positions in already aligned sequences, insert dummy characters into the internode sequence. Dummy characters are treated such that there is no penalty of any sort (mismatch or gap) associated with them.

Some ClustalW wrinkles:

At internodes, Clustalw infers consensus-type sequences to represent the internode for the next step in the progressive alignment.

These inferred internode sequences are constructed by downweighting sequences that are closely related to one another. In this way, a large group of closely related sequences will not dominate construction of the consensus-type sequence over a solitary distantly related sequence.

New gaps formed during the progressive alignment procedure are penalized less if an existing gap exists already at the position.

New gaps formed near but not at the position of an existing gap are penalized heavily.

PAM/BLOSUM weight matrices vary according to the length of the branch used to guide the alignment.

There are reduced gap penalties in hydrophilic sequence stretches.

Problems of multiple sequence alignment methods that are based on dynamic programming

1. Problem of parsimony and weighting.

A. Relative weights of different events

B. Relative weights of different sequence positions (addressed in ClustalW).

C. More general criticisms of parsimony approachs (*i.e.*, Where's the model? What's the biology?)

2. Present strategies consider only a very few possible multiple sequence alignments when constructing phylogeny. Ideally, all alignments would be considered in relation to their probability.

3. The best way to detect or measure relationship between a pair of sequences is to use the distribution of all possible pairwise alignments, not just the best single alignment.

4. Inability to deal with more exotic evolutionary events such as inversions and gene conversion in a statistically valid, computationally feasible manner.

How can global alignment methods be evaluated?

In most cases, the "true" alignment cannot be known with certainty.

1. Simulation of data sets

Problem: What model of insertion and deletion to use?

2. Alignment of functional motifs

(approach taken by McClure et al. 1994)

Potential Problem: Can we be sure about the alignment of the motifs?

3. Correspondence of alignments with protein structure in cases where protein structures are experimentally determined

Is structure the "gold standard" of sequence alignment?

Potential Misconception: Known structures can add information to the alignment inference but they cannot solve the alignment problem.

4. Gestalt and aesthetics

Multiple sequence alignment references

H. Carillo and D. Lipman. 1988. The multiple sequence alignment problem in biology. SIAM Journal on Applied Mathematics. 48(5): 1073–1082.

Hein, J. 1990. A unified approach to alignment and phylogenies. Pp. 626–645 in R.F. Doolittle, ed. Methods in Enzymology, Vol. 183. Academic Press, San Diego. (**Describes TreeAlign, a program that combines sequence alignment and phylogeny inference**)

Lipman DJ, Altschul SF, Kececioglu JD. 1989. A tool for multiple sequence alignment. Proc. Natl. Acad. Sci. USA 86:4412-4415.

McClure MA, Vasi TK, and WM Fitch. Comparative analysis of multiple-sequence alignment methods. Mol Biol Evol 11(4):571-592.

Sankoff D. 1975. Minimal mutation trees of sequences. SIAM J. Appl. Math. 78:35–42. (**The three papers by Sankoff and collaborators that are listed were way ahead of their time. These authors understood that sequence alignment and phylogeny inference should be done simultaneously**).

Sankoff, D., C. Morel, and R.J. Cedergren. 1973. Evolution of 5S RNA and the nonrandomness of base replacement. Nature New Biol. 245:232–234.

Sankoff, D., R.J. Cedergren, and G. Lapalme. 1976. Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA. J. Mol. Evol. 7:133–149.

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence–weighting, position–specific gap penalties and weight matrix choice. Nucleic Acids Research 22:4673–4680. (**The latest in a series of papers by Desmond Higgins and collaborators on the CLUSTAL alignment software**).