Global Multiple Sequence Alignment

Stephen F. Altschul

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health

Why Multiple Alignment?

Optimal pairwise alignments may be inconsistent.

> In other words, letter “A” from sequence 1 may align with “B” from sequence 2 and “C” from sequence 3, but “B” from sequence 2 does not align with “C” from sequence 3.

Ambiguities in how best to align two sequences may be resolved when other sequences are available.

Patterns of conservation may become apparent only when many sequences are aligned.

What Are We Trying To Optimize?

```
Human:     VHLTPEEKSAVTALW----GKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK…
Lemur:     TFLTPEENGHVTSLW----GKVNVEKVGGEALGRLLVVYPWTQRFFESFGDLSSPDAIMGNPK…
Goldfish:  VEWTDAERSAIIGLW----GKLNPDELGPQALARCLIVYPWTQRYFATFGNLSSPAAIMGNPK…
Bloodworm: MGLSAAQRQVVASTWKDIAGSDNGAGVGKECFTKFLSAHHDIAAVF-GFSGAS-------DPG…
Soybean:   VAFTEKQDALVSSSFE--AFKANIPQYSVVFYTSILEKAPAAKDLFSFLANGVDPT----NPK…
```

How should one define substitution scores?

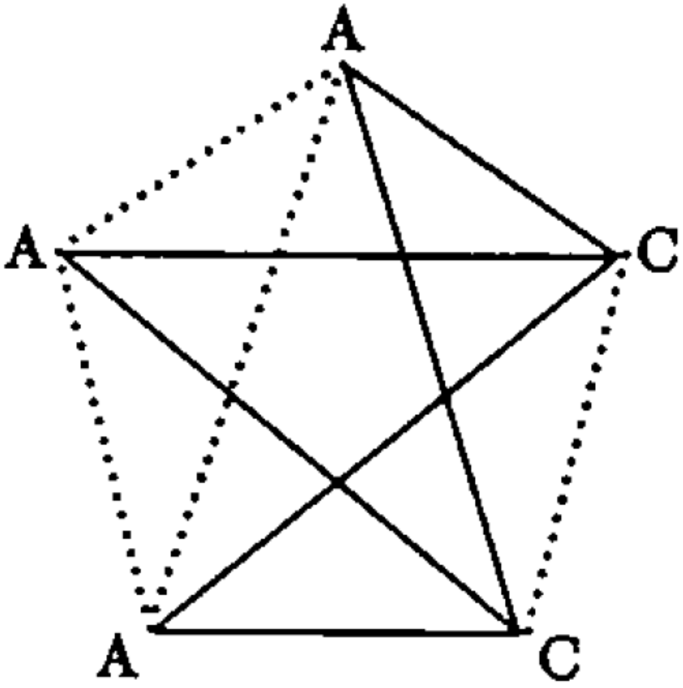
How should one define gaps and gap scores?

Should one take account of an evolutionary tree relating the sequences? (Should one construct the tree? How?)

Should one take account of sequence correlations in the absence of a tree? (How should this be done?)

Multiple Alignment Substitution Scores

a) Sum-of-the-pairs or SP-scores



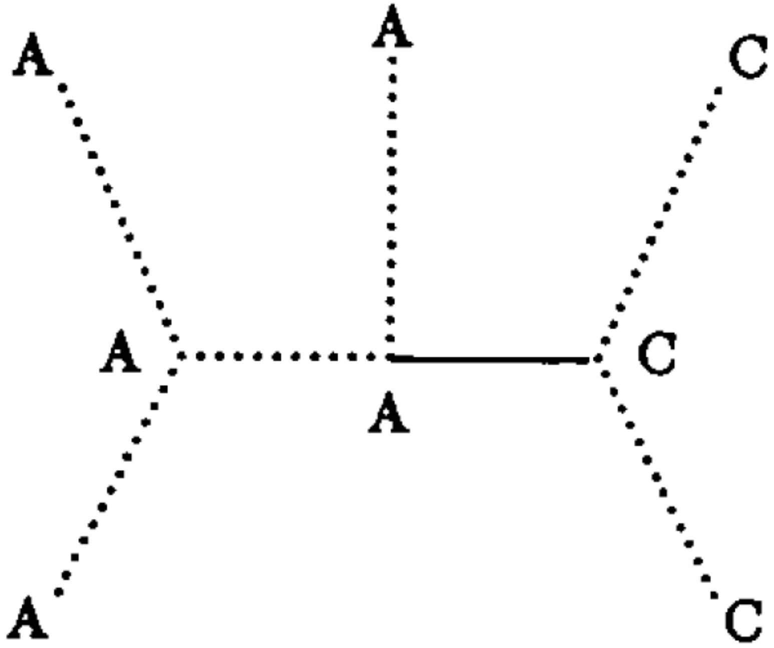
4 matches; 6 mismatches

Murata, M., *et al.* (1985) "Simultaneous comparison of three protein sequences." *Proc. Natl. Acad. Sci. USA* **82**:3073-3077.

Bacon, D.J. & Anderson, W.F. (1986) "Multiple sequence alignment." *J. Mol. Biol.* **191**:153-161.

Multiple Alignment Substitution Scores

b) Tree scores

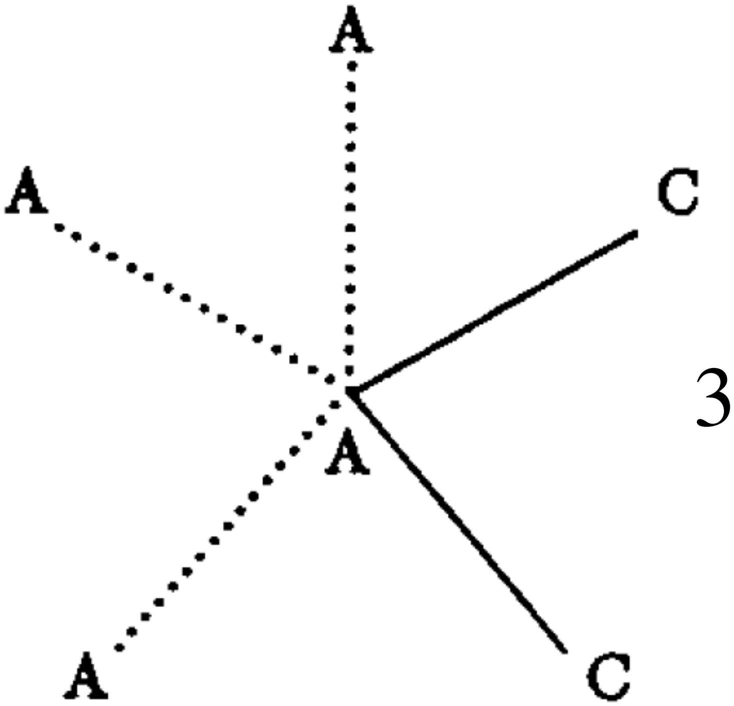


6 matches; 1 mismatch

Sankoff, D. (1975) “Minimal mutation trees of sequences.” *SIAM J. Appl. Math.* **28**:35-42.

Multiple Alignment Substitution Scores

c) Star or consensus scores



3 matches; 2 mismatches

Multiple Alignment Substitution Scores

d) Entropy-based scores

A
A
A
C
C

$$\log(4) - 0.6\log(0.6) - 0.4\log(0.4)$$

$$= 1.03 \text{ bits}$$

Schneider, T.S., *et al.* (1986) “Information content of binding sites on nucleotide sequences.” *J. Mol. Biol.* **188**:415-431.

Multiple Alignment Substitution Scores

e) Log-odds scores

$$S(\vec{x}) = \log \frac{Q(\vec{x})}{P(\vec{x})}$$

"Bayesian Integral Log-odds" or "BILD" scores

The construction of column scores from Dirichlet mixture priors

$$Q(\vec{x}) = \sum_{i=1}^{M} m_i \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + c)} \prod_j \frac{\Gamma(\alpha_{i,j} + c_j)}{\Gamma(\alpha_{i,j})} \qquad P(\vec{x}) = \prod_k p_{x_k}$$

where $\vec{c}$ is the amino acid count vector implied by $\vec{x}$

Assuming uniform Dirichlet priors, $S(\text{"AAACC"}) = \log(1.83) = \quad 0.87$ bits

$S(\text{"AAACT"}) = \log(0.91) = -0.13$ bits

Altschul, S.F., *et al.* (2010) "The construction and use of log-odds substitution scores for multiple sequence alignment." *PLoS Comput. Biol.* **6**:e1000852.

Multiple Alignment Gap Scores

Gap scores should, in general, be defined consistently with substitution scores.

For example, if “SP” substitution scores are used, gap scores should also be defined as the sum of gap scores for the implied pairwise alignments.

Following this prescription completely rigorously for affine gap scores entails unacceptable algorithmic complications, which can be avoided by a slight modification of one’s definition of gap score.

Altschul, S.F. (1989) “Gap costs for multiple sequence alignment.” *J. Theor. Biol.* **138**:297-309.

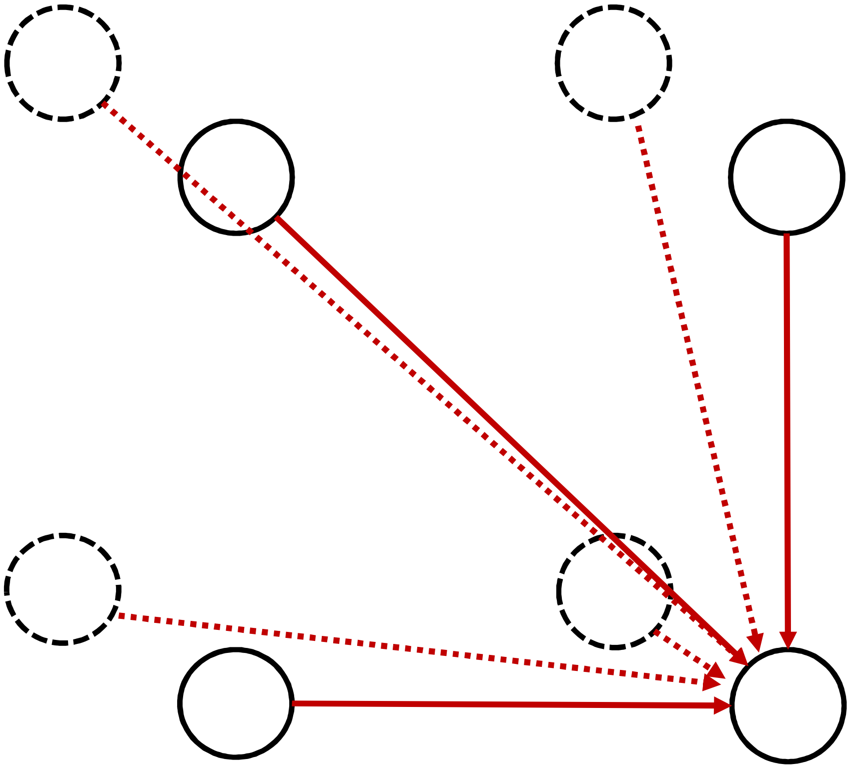
Multiple Alignment Algorithms

Some multiple alignment algorithms assume or require a particular type of score, while others may permit a variety of scores.

Some multiple alignment algorithms are defined purely procedurally, without reference to any explicit objective function they are trying to optimize.

By most definitions of the problem, multiple alignment is hard, so most practical algorithms that have an explicit objective function are heuristic.

Generalization of Dynamic Programming

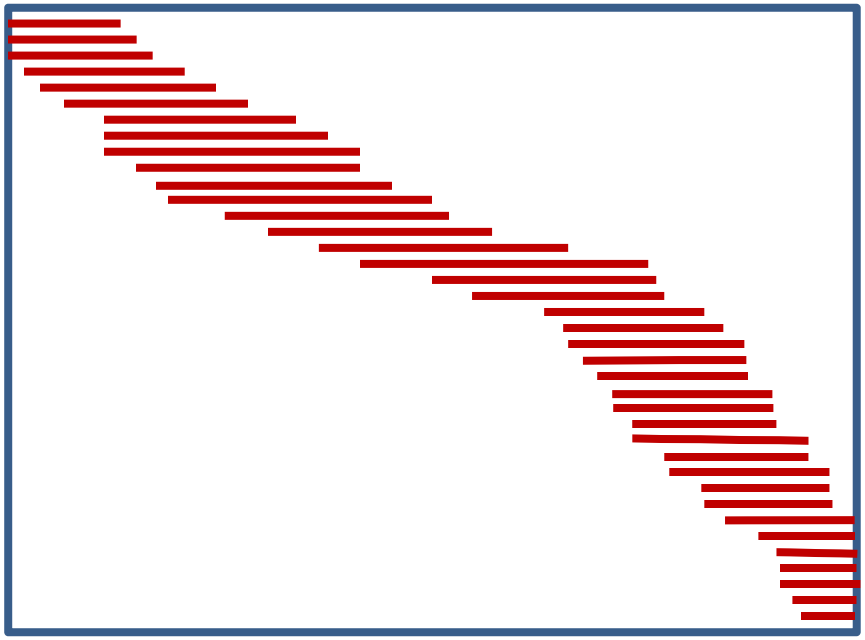


For k sequences, each of length n, there are n^k nodes, and most nodes require optimizing scores among $2^k - 1$ incoming edges. This yields a time complexity of $O\left((2n)^k\right)$.

Rigorous dynamic programming is feasible for at most three or four sequences of typical length.

Adding affine gap costs introduces further complications;
see: Altschul, S.F. (1989) *J. Theor Biol.* **138**:297-309.

Speeding Up Dynamic Programming



For each pair of sequences, find those nodes through which pairwise alignments with score within ε of the optimum pairwise score may pass.

When performing multidimensional dynamic programming, consider only nodes whose projections onto each pair of sequences fall within the permitted regions.

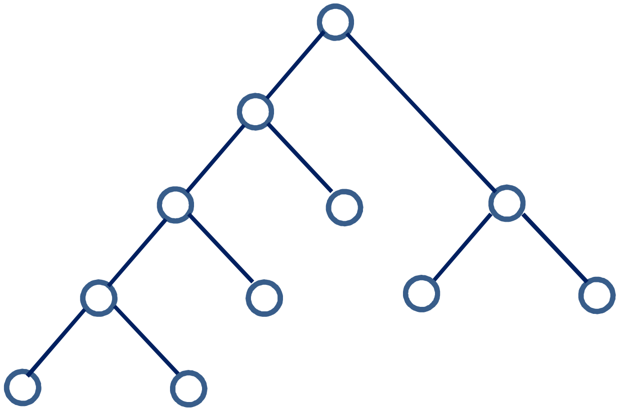
It is possible to choose ε for the various pairs so that an optimal *k*-dimensional alignment is guaranteed to project within the allowed 2-dimensional regions. However, this approach may also be used to construct a heuristic algorithm, extending dynamic programming to as many as seven or eight sequences.

Carrillo, H. & Lipman, D. (1988) *SIAM J. Appl. Math.* **48**:1073-1082.

Lipman, D.J. *et al*. (1989) *Proc. Natl. Acad. Sci. USA* **86**:4412-4415.

Progressive Alignment

An alternative, heuristic method for multiple alignment is the "progressive" approach, in which pairs of sequences, or of fixed alignments, are progressively aligned two one another using a standard pairwise alignment algorithm. An alignment of two or more sequences, once fixed, is not changed when additional sequences are added. Given k sequences of length n, and assuming the length of the overall alignments do not grow unduly, a total of $k - 1$ pairwise alignments are performed, each requiring n^2 time, yielding a time complexity of $O((k-1)n^2)$.



A guide tree for six sequences

A common practice is to align sequences or groups of sequences in the order dictated by a rooted "guide tree", from the leaves upward, with individual sequences assigned to the leaves. The idea is usually to align the most closely related sequences first. The guide tree is sometimes constructed from a set of pairwise distances or similarities, but calculating these distances can require $O(k^2n^2)$ time, becoming the rate limiting step.

Progressive Alignment Programs

Many additional ideas go into the construction of practical multiple alignment programs, and the problem is by no means solved. Here are some of the multiple alignment programs most widely used today:

Thompson, J.D., Higgins, D.G. & Gibson, T. J. (1994) "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." *Nucl. Acids Res.* **22**: 4673-4680.

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002) "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." *Nucl. Acids Res.* **30**:3059-3066.

Edgar, R.C. (2004) "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucl. Acids Res.* **32**:1792-1797.