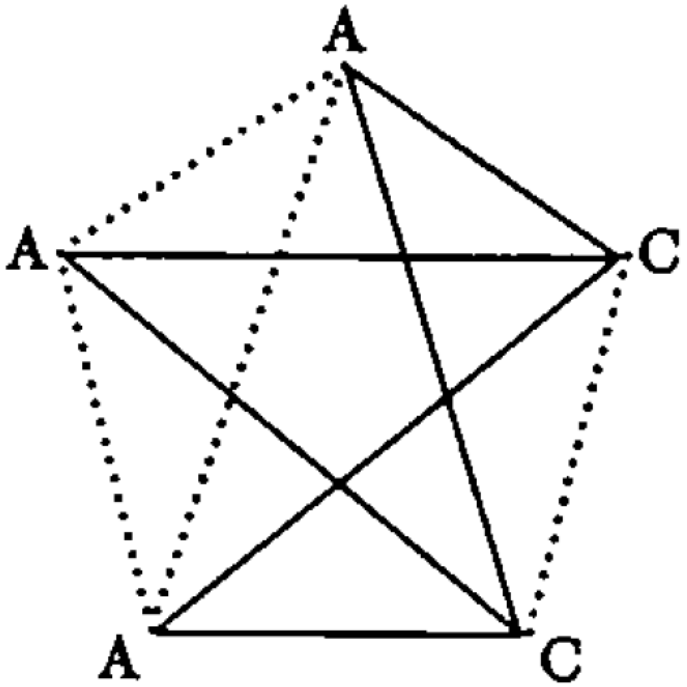
Local Multiple Sequence Alignment

Stephen F. Altschul

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

Multiple Alignment Substitution Scores

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



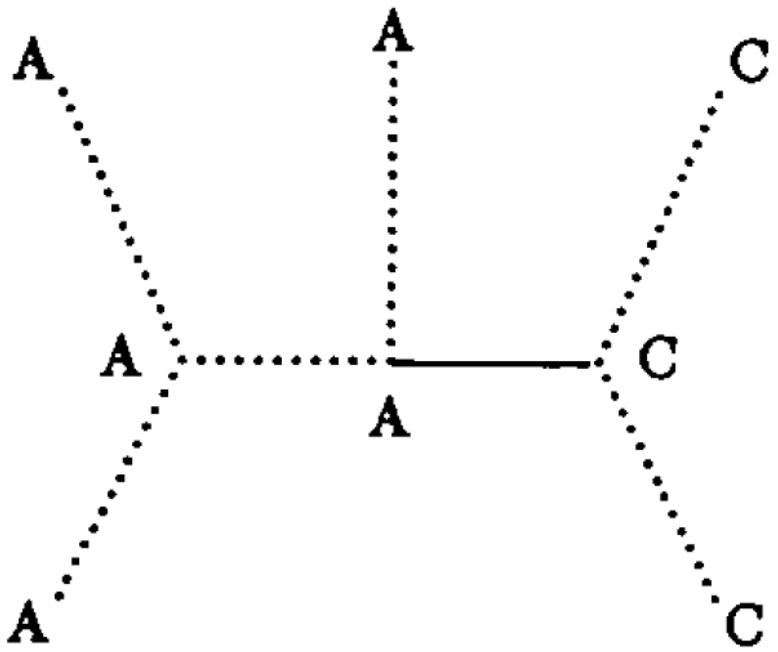
4 matches; 6 mismatches

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

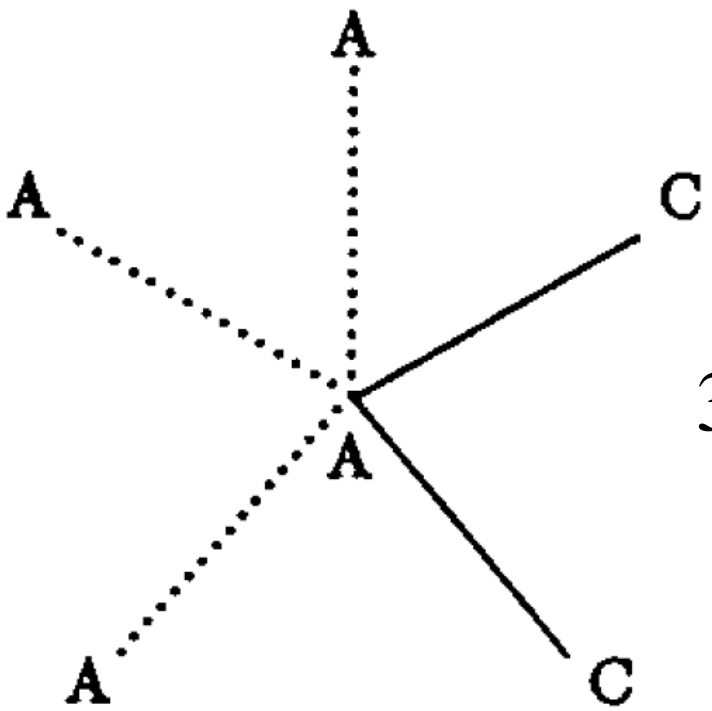


6 matches; 1 mismatch

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

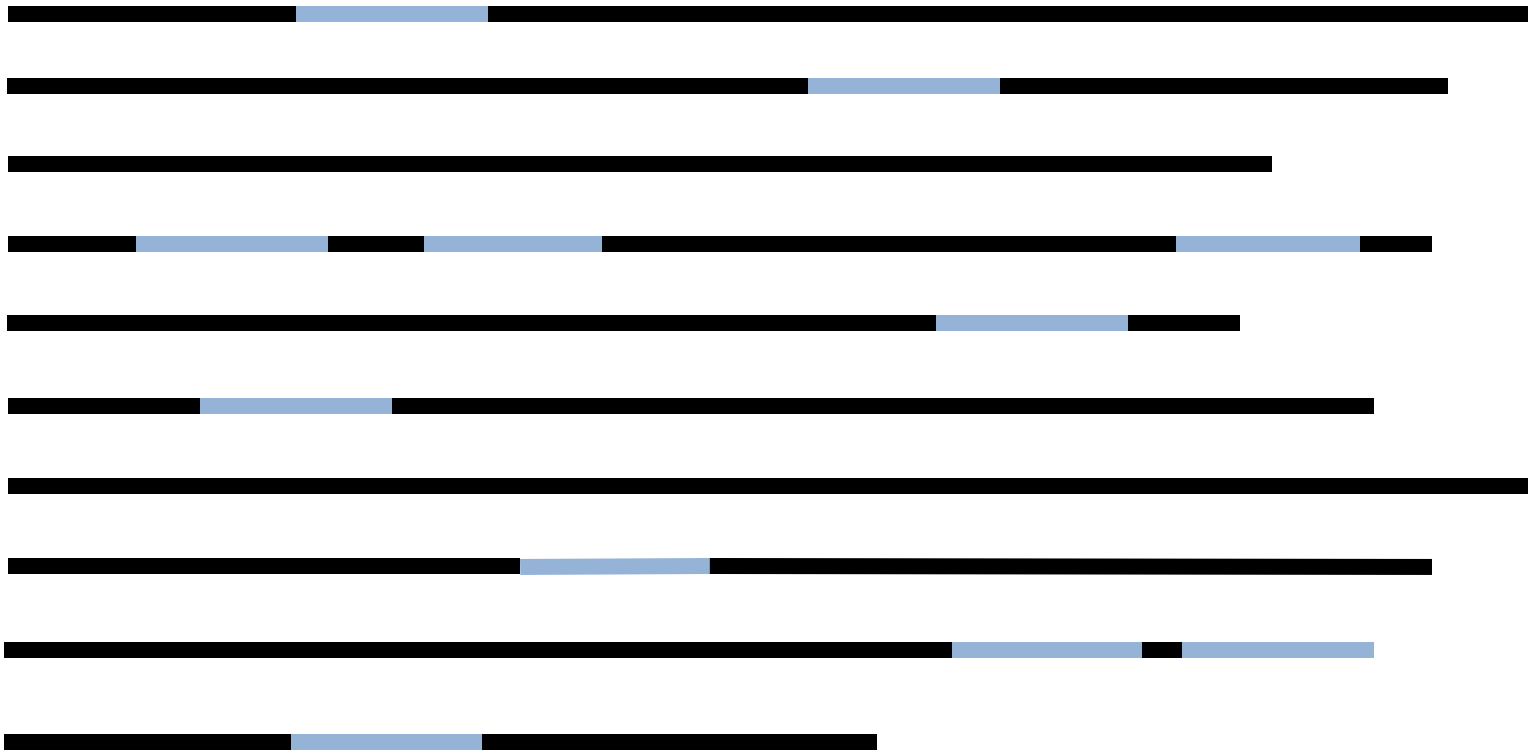
Gap scores should, in general, be defined consistently with substitution scores.

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.

Local Multiple Alignment: The Problem



Neither pattern nor locations known

Local Multiple Alignment

Desiderata for an ideal local multiple alignment algorithm:

- Employs an appropriate measure of alignment quality
- Measure can reflect known amino acid relationships (proteins)
- Width of pattern not unduly constrained
- Width need not be specified *a priori*
- Pattern may be missing or present in multiple copies
- Alignment of segments may contain gaps
- Output is independent of order of input sequences
- Algorithm can find multiple distinct patterns
- Algorithm is deterministic
- Algorithm is rigorous optimization procedure
- Time complexity is linear in number of sequences

Approaches to Local Multiple Alignment

Consensus Word Methods
Template Methods
Progressive Alignment Methods
Pairwise Consistency Methods
Statistical Methods

We will consider only approaches that do not allow gaps

Notation

Number of sequences:	N
Average length of sequences:	L
Size of alphabet:	A
Specified width of pattern:	W

Consensus Word Methods

Find a **consensus word** that is "close" to a word in each, or a large number of the input sequences. This may be thought of as finding an optimal star alignment.

Algorithmic outline

For each word of fixed width W, define a *neighborhood* of B adjacent words, each with an associated score.

Example: For DNA, words could be 8-tuples, and the neighborhood of X all those words with no more than 2 mismatches with X.

Given each sequence S, for each word X in S, update the "best match to S" for all neighbors of X.

Time complexity: $NLB + A^W$. Space complexity: A^W.

Advantages

For fixed word and neighborhood size, time complexity is linear in input data. Given definition of problem, algorithm is rigorous.

Disadvantages

Pattern length predefined. Fairly severe restrictions on pattern length and neighborhood size. No scores for words outside neighborhood.

Verdict: May be OK for some DNA applications; of very little use for proteins.

Queen, C.M. *et al.* (1982) *Nucl. Acids Res.* **10**:449-456.

Galas, D.J., *et al.* (1985) *J. Mol. Biol.* **186**:117-128.

Waterman, M.S., *et al.* (1984) *Bull. Math. Biol.* **46**:515-527.

Staden, R. (1989) *Comput. Appl. Molec. Biol.* **5**:293-298.

Template Methods

Search for a set of **templates** within each input sequence.

Algorithmic outline

Define a set of templates of total size B.

Example: For protein sequence comparison, a template could be "V*C**D", where '*' is a wild card.

Compare each template to all input sequences, updating a score for the template whenever a match is found.

Time complexity: NLB. Space complexity: B.

Comment

This is basically an inversion of the consensus word methods, but with processing done one template at a time, rather than one sequence at a time.

Advantages and Disadvantages

Essentially the same as those for the consensus word methods.

Verdict: More flexible than consensus word methods, but with similar major limitations for protein comparison.

Sobel, E. & Martinez, H. (1986) *Nucl. Acids Res.* **14**:363-374.

Posfai, J., *et al.* (1989) *Nucl. Acids Res.* **17**:2421-2435.

Smith, H.O., *et al.* (1990) *Proc. Natl. Acad. Sci. USA* **87**:826-830.

Leung, M.Y., *et al.* (1991) *J. Mol. Biol.* **221**:1367-1378.

Progressive Alignment Methods

Build up local multiple alignments of fixed width in a progressive manner.

Algorithmic outline

Select a fixed pattern width W. Compare all segments of this width in the first sequence to all such segments in the second, using an arbitrary scoring system. Retain the B best pairs. Compare these to all segments in the third sequence, *etc.*

Time complexity: $NLBW$

Variations

Retain the best multiple alignment for each segment from the first sequence.

Advantages

No significant restriction on pattern width. For fixed B and W, linear time in length of input data. Can use arbitrary score function.

Disadvantages

Heuristic: optimal solution not guaranteed. Dependent on sequence order. Parameter B may need to be very large to yield good results.

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

Stormo, G.D. & Hartzell, G.W. III (1989) *Proc. Natl. Acad. Sci. USA* **86**:1183-1187.

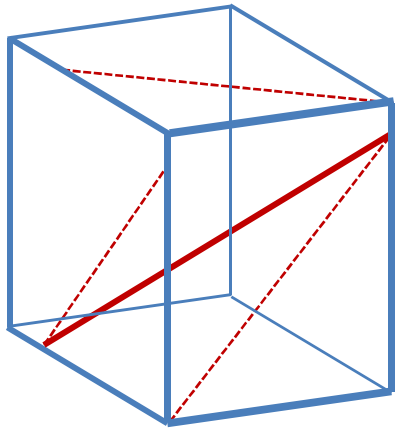
Hertz, G.Z., *et al.* (1990) *Comput. Appl. Molec. Biol.* **6**:81-92.

Pairwise Consistency Methods

Compare all pairs of sequences. Seek consistency among aligned letters, or diagonals.

Algorithmic outline (Schuler *et al.*)

Execute ungapped Smith-Waterman algorithm on all pairs of sequences. Mark all *diagonals* containing segment pairs that exceed a threshold score H. Build up "high-dimensional" diagonals, all (or almost) all of whose pairwise projections have been marked. Search any such high-dimensional diagonals for high-scoring ungapped local multiple alignments.



Time complexity: $N^2L^2 + f(H)$

Advantages

No predefined pattern width required. Can find multiple distinct patterns. Can use arbitrary scoring system. Need not include all sequences. Rigorous optimization procedure, given constraint on pairwise projection scores.

Disadvantages

Quadratic time in input length. Space and time complexity balloon for small H.

Comment

Good for a moderate number of sequences. Implemented in interactive "MACAW" program.

Schuler, G.D., *et al.* (1991) *Proteins* **9**:180-190.

Vingron, M. & Argos, P. (1991) *J. Mol. Biol.* **218**:33-43.

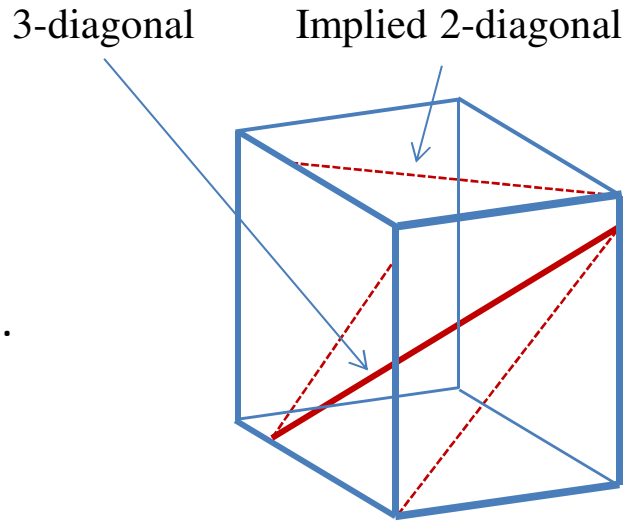
The Number of Consistent Diagonals

Consider adding one sequence of length L at a time.

Once one has added N sequences, the number of N-diagonals is approximately NL^{N-1}.

The number of 2-diagonals implied by an N-diagonal is $\binom{N}{2}$.

Assume the probability of a 2-diagonal being marked is p, and that all 2-diagonals are independent.



Approximate no. of "random" consistent N-diagonals, assuming $L = 1000$, and $p = 0.1$.

N	N-diagonals	Implied 2-diagonals	Consistent N-diagonals
2	2×10^3	1	200
3	3×10^6	3	3,000
4	4×10^9	6	4,000
5	5×10^{12}	10	500
6	6×10^{15}	15	6
7	7×10^{18}	21	0.007

Statistical Methods

Local multiple sequence alignment can be view as an optimization problem in a rough, high-dimensional space.

One may approach this classic problem with the deterministic *expectation-maximization (EM)* method (Dempster, *et al*., 1977).

Alternatively, one may apply one of the related stochastic methods of *simulated annealing* (Metropolis, *et al*., 1953) or *Gibbs sampling* (Geman & Geman, 1984).

Applied to local multiple sequence alignment, these approaches alternate between refining a provisional pattern, based upon its assumed locations within the sequences, and updating these locations, given the pattern.

Metropolis, N., *et al.* (1953) *J. Chem. Phys.* **21**:1087-1092. Dempster, A.P., *et al.* (1977) *J. Roy. Stat. Soc. B* **39**:1-38.

Geman, S. & Geman, D. (1984) *IEEE Trans. Pattern Analysis and Machine Intelligence* **6**:721-741.

Lawrence, C.E. & Reilly, A.A. (1990) *Proteins* **7**:41-51. Cardon, L.R. & Stormo, G.D. (1992) *J. Mol. Biol.* **223**:159-170.

Lawrence, C.E., *et al.* (1993) *Science* **262**:208-214.