Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 8777–8781, October 1993
Biochemistry

Weighting in sequence space: A comparison of methods in terms of generalized sequences

(alignment/profiles/correcting for correlation/sequence weighting)

MARTIN VINGRON* AND PETER R. SIBBALD†

*Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113; and †EMBL Data Library, European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10 22 09, W-6900 Heidelberg, Germany

Communicated by Manfred Eigen, May 17, 1993

ABSTRACT **Four methods for weighting aligned biological sequences have recently appeared that differ mathematically, philosophically, and in their results. Thus, while there is consensus about the need to weight sequences, the method to use is contentious. A geometric analysis based on a continuous sequence space is presented that provides a common framework in which to compare the methods. It is concluded that there are two "best" methods. When the sequences are known to be phylogenetically related and a tree can be generated without introducing excessive stress into the data, the method of Altschul *et al.* [Altschul, S. F., Carroll, R. J. & Lipman, D. J. (1989) *J. Mol. Biol.* 207, 647–653] is appropriate. When the sequences are not known to be phylogenetically related or a tree cannot be produced without unduly distorting the distances between the sequences, a modification of the method of Sibbald and Argos [Sibbald, P. R. & Argos, P. (1990) *J. Mol. Biol.* 216, 813–818] is preferable.**

Correlated observations can complicate analysis of biological data sets (1, 2). When independence cannot be assumed, it is erroneous to proceed as if all the data were equally informative. In the problem of multiple sequence alignment this is important because alignments frequently contain very similar (even duplicated) sequences; these can bias the construction of the alignment itself (3–5) or make some trends (merely due to nonrandom sampling) appear strong. Equally deleterious is the "swamping" of interesting but rare data. This is a problem in any analysis of multiply aligned DNA or amino acid sequences; searches of data bases with multiple alignments (6) are sensitive to the frequency with which aligned sequences occur. Similar problems arise in predicting protein secondary structure from multiple alignments (7). Scores averaged over alignment columns are vulnerable to over- or under-representation of certain sequences. A correction may also be applied when phylogenetic relatedness or structural constraint restricts the range of diversity within a subset of sequences. A remedy is to assign weights to the sequences in an alignment before calculating any average value; this is termed *sequence weighting*.

Increased sequence data and fast multiple-alignment programs (e.g., refs. 8–12) have aggravated the problem. It might be argued that a profile created by an expert with reasonable, representative sequences would eliminate any need for weighting. Such an approach is not objective and loses information, and as data sets grow, automation is essential. Consequently, there has been interest in methods to weight aligned sequences (10, 13–15).

Most discrepancies between the methods stem from divergent definitions of the problem and lack of agreement concerning "correct" behavior. The methods often provide different results and are based on different reasoning, but all claim to be solving approximately the same problem. In each method a weight for each sequence in the alignment is derived, greater weights indicating greater importance. The Altschul, Carroll, and Lipman (1989) (ACL) method (14) differs by down-weighting distant sequences as contributing less information to the point of interest, whereas the other methods up-weight outliers as contributing more information about the diversity of the data set. Inaccuracies in some methods have come to light. Because errors tend to propagate (e.g., alignments are typically used to calculate hydropathy or similar indices and influence work based on them), we compare the methods and describe known failings.

Continuous Sequence Space

Attributing weights to aligned sequences is a one-dimensional and thus crude scaling. Sophisticated relationships sought by evolutionary trees (1, 16, 17), statistical geometry (18), or multidimensional scaling (19, 20) are simplified drastically. Discrete sequence space (18) can be thought of as the set of all possible sequences of a given length. In a simple version, one of two letters—e.g., R and Y—is allowed at every position of a sequence. For length $L = 2$ all sequences can be visualized as corners of a square (Fig. 1*a*). With increased L the sequence space can be visualized as a hyper-cube of dimension L (imagine a triangle representing three letters).

We define a profile as a matrix where columns contain the distribution vectors of letters in corresponding alignment columns (ref. 10; this is slightly different use of the word "profile" than in ref. 6). If a three-sequence alignment column reads R R Y, this is represented as ⅔ R and ⅓ Y. Independent of the number of aligned sequences, profiles have as many rows as there are letters in the alphabet. Profile positions are distribution vectors falling (excepting completely conserved positions) into the "empty space" between corners (Fig. 1*b*) and, therefore, cannot be represented in discrete sequence space.

Discrete sequence space can be extended to a continuous sequence space. We term the elements *generalized sequences*. A generalized sequence is a matrix with as many rows as letters in the alphabet, where the sum over each column is 1. The profile of an alignment is a generalized sequence corresponding to this alignment, but while profiles are generated from alignments in a very specific way, generalized sequences may be generated by other means (*vide infra*).

In discrete sequence space, the Hamming distance (for two sequences of equal length) counts the mismatches between characters. On biological sequences a similarity or dissimilarity matrix on the letters of the alphabet (e.g., the amino

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: Methods are indicated as follows: ACL, Altschul, Carroll, and Lipman (1989); SS, Sander and Schneider (1991); VA, Vingron and Argos (1989); VOR, Sibbald and Argos (1990) "Voronoi"; mVOR, "modified Voronoi."

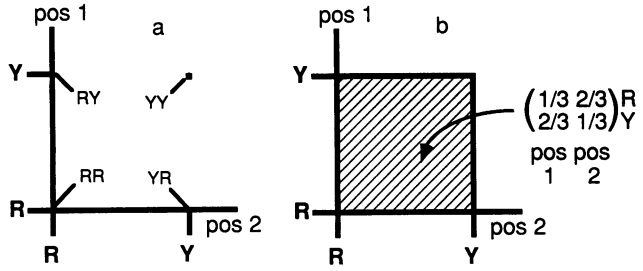

FIG. 1. Discrete and continuous sequence space. (*a*) The discrete space of sequences of length 2 over the alphabet R,Y. (*b*) A generalized sequence in its location in continuous sequence space. pos, Position.

acids) is used and defines the (dis)similarity between two sequences as the sum of the positional (dis)similarity values. The difference between a similarity and a dissimilarity lies in its interpretation. A similarity matrix attributes high values to "desirable" pairs of letters, which in a dissimilarity matrix would receive a low value—e.g., 0. There is no need to limit our discussion to either similarities or dissimilarities, and we write "(dis)similarity" to indicate this fact. Euclidian distances on two generalized sequences have been used (4) but with profiles derived mostly from alignments, it has become customary (6, 10) to use the average (dis)similarity between the sequences involved instead. Note the differences between average dissimilarity and distance: an average dissimilarity between two identical alignments need not be 0 (even if the main diagonal of the matrix is all 0s).

We define (dis)similarity for generalized sequences such that all discrete cases correspond to a normal (dis)similarity measure. Consider two positions (columns) p and q of two generalized sequences of equal length. Given a matrix M‡ on the letters we define the (dis)similarity $d_M(p, q)$ of the two distribution vectors p and q as the inner product $p \cdot Mq = \Sigma_{i,j} p_i q_j m_{ij}$. When M is symmetric, this equals $q \cdot Mp$. To compare two generalized sequences these positional inner products are summed over the sequences. When two positions correspond each to a single letter (unit vectors when translated into profiles), the inner product reduces to the corresponding entry of M. Therefore, when applied to a generalized sequence describing a conventional one, positional values are summed over the whole sequence, and the general (dis)similarity measure reduces to the normal one to compare two sequences. The gap may be treated as an additional letter in the alphabet.

Let the alignment be s and count the different letters in position i by summing unit vectors, each having a 1 for one sequence: $\Sigma_k e_{s_{ik}}$, where k goes from 1 to the number of sequences N. The resulting vector has the number of occurrences of the first letter in alignment column i in its first component etc. The profile column $p(s_i)$ is therefore

$$p(s_i) = 1/N \sum_{k=1}^{N} e_{s_{ik}}. \quad [1]$$

Geometrically, $p(s_i)$ describes the center of gravity (4) of sequences at position i.

‡Special symbols are as follows: N, number of sequences; L, length of sequences = length of alignment; $\mathcal{A}$, alphabet; $|\mathcal{A}|$, size of alphabet; $s = (s_{ik})_{i,k}$, alignment, $1 \le i \le L$, $1 \le k \le N$; s_i, column of alignment s; s_k, kth sequence; $\mathbf{p} = (p_{ik})_{i,k}$, $1 \le i \le L$, $1 \le k \le |\mathcal{A}|$; $p(s_i)$, column i of profile $\mathbf{p}(s)$; e_a, $a \in \mathcal{A}$, unit vector in $\{0, 1\}^{|\mathcal{A}|}$ with 1 at position of letter a; $M = (m_{ij})_{i,j}$, (dis)similarity matrix on $\mathcal{A}$, $1 \le i,j \le |\mathcal{A}|$; $D = (d_{ij})_{i,j}$, (dis)similarity matrix on the set of sequences, $1 \le i,j \le N$.

Given an alignment, between any two sequences we calculate a (dis)similarity and obtain an $N \times N$ matrix of (dis)similarities $D = (d_{ij})$, $1 \le i,j \le N$. From the definition of the (dis)similarity between generalized sequences as an inner product, it is easy to see *the (dis)similarity between any one of the sequences, say* k, *and the profile of the alignment is just the average (dis)similarity between sequence* k *and the others*. This is also a consequence of the next case, the weighted profile.

For N sequences, assume a vector $w = (w_1, \ldots, w_N)$ of weights normalized to sum to 1. Define the weighted profile $\mathbf{p}_w(s)$ for a set of aligned sequences as the generalized sequence with the following columns:

$$p_w(s_i) = \sum_k w_k e_{s_{ik}}. \quad [2]$$

This weighted profile differs from the simple profile because it is a weighted average of single sequences. Let M be a (dis)similarity matrix on the letters of the alphabet. To limit the depth of the indices we write $m(a, b)$ instead of m_{ab}. The (dis)similarity between sequences k and l, d_{kl}, equals $\Sigma_{i=1 \ldots L} m(s_{ik}, s_{il})$. *The (dis)similarity between a sequence* k *and a weighted profile turns out to be the weighted average of the individual (dis)similarities*:

$$\sum_{l=1}^{N} w_l d_{kl} = \sum_{l=1}^{N} w_l \sum_{i=1}^{L} m(s_{ik}, s_{il}) = \sum_{i=1}^{L} e_{s_{ik}} \cdot M \left(\sum_{l=1}^{N} w_l e_{s_{il}} \right)$$

$$= \sum_{i=1}^{L} e_{s_{ik}} \cdot M \mathbf{p}_w(s_i) = d[s_k, \mathbf{p}_w(s)]. \quad [3]$$

This means the matrix D applied to the weight vector w coincides with the vector of (dis)similarities between the sequences and the w-weighted profile:

$$Dw = \{d[s_k, \mathbf{p}_w(s)]\} k = 1 \ldots N.$$

Weighting: A Comparison of the Methods in the Common Framework

One objective of sequence weighting is to prevent several similar sequences outweighing a few "atypical" ones. Vingron and Argos (1989) (VA) (10) simply observed atypical sequences in an alignment differ from the others at many positions and proposed a (normalized) count of the mismatches between one sequence and all others as the weight of that sequence: an "outsider" receives higher weight. In terms of generalized sequences, the average count of all mismatches between a sequence and an alignment is just the average dissimilarity between that sequence and the others—i.e., the dissimilarity between a sequence in an alignment and the alignment profile. *Thus, the weights are the dissimilarities of the sequences from their center of gravity*. The compact description derived above allows a summary of the N equations $w_k = \Sigma_i d_{ik}$, $k = 1 \ldots N$, as

$$w = D \cdot (1, 1, \ldots, 1) = D \cdot \mathbf{1},$$

where 1 is an abbreviation for the vector containing all 1s, and D denotes the matrix of dissimilarities d_{ij}.

Sander and Schneider (1991) (SS) (15) argued that under those weights a new profile and new distances should be calculated. They proposed a self-consistent set of weights placing the centroid such that the dissimilarities from the sequences to the centroid equal (up to a factor) the weights determining the centroid. Using matrix notation, they demand that

$$\lambda w = Dw.$$

This specifies an eigenvector w with eigenvalue λ. A simple numerical method for finding an eigenvector by repeated application of the matrix to a starting vector can be used. For a matrix A the process

$$w^{(k)} \leftarrow Aw^{(k-1)}$$

converges to eigenvector w with the largest eigenvalue. Because distance matrix D has all entries ≥ 0, a theorem due to Perron and Frobenius on nonnegative matrices (21) guarantees good behavior of this method, excepting degenerate cases where the iteration oscillates. For distance matrices a remedy is to add the identity matrix to A, permitting the iteration to converge to the eigenvector (22).

The VA method increases the weight of a sequence far from the (unweighted) profile and moves the centroid away from nearby sequences and toward distant ones. The objective may be interpreted so as to make the dissimilarities between all sequences and centroid equal to some value, e.g., 1:

$$[d(s_k, \mathbf{p}_w(\mathbf{s})]_{k=1 \ldots N} = Dw = \mathbf{1}. \quad [4]$$

If D is invertible, the weights can be calculated as $w = D^{-1}\mathbf{1}$ (w is then normalized). This is very similar to the description of the ACL method (14), which uses a variance–covariance matrix A and calculates weights as $w = A^{-1}\mathbf{1}$.

The "Voronoi" Sibbald and Argos (1990) (VOR) method (13) takes into account distances to the other sequences but not the centroid. To each sequence its Voronoi-cell is attributed—i.e., the set of points closest to this one sequence. The more isolated a sequence is, the greater is its volume. This volume is calculated by a Monte–Carlo algorithm building random sequences from amino acids occurring at each alignment position. If such a random sequence is closest to n sequences, the weight of each of the n sequences is incremented by $1/n$. Although conceived for cases where n sequences are identical, it applies in other cases. This can lead to "incorrect" weights: consider the example alignment below. Intuitively AA and BB must receive equal weight (1/2), and the two copies of AA share this 1/2 such that the relative weights are 1:1:2 ("correctness" is discussed below). However, applying the sampling scheme described above one obtains the following result:

	Generated random sequences				
	AA	AB	BA	BB	Sum
AA	1/2	1/3	1/3	0	7/6
AA	1/2	1/3	1/3	0	7/6
BB	0	1/3	1/3	1	10/6

The discrepancy is due to one-half (namely, AB and BA) of all randomly generated sequences being equidistant to two different others (AA, BB). For longer sequences the likelihood for this event is negligible and of little practical relevance. A minor modification solves the problem: instead of generating discrete random sequences one can choose random generalized sequences to estimate the volume of the Voronoi cells in continuous sequence space. To ensure randomly generated sequences, uniformly distributed in sequence space we generate each column by normalizing a set of independent, exponentially distributed random numbers (ref. 23, problem 1.2.6). Even in contrived cases like the given one, there is 0 probability of a generalized sequence being equidistant from two distinct ones. The above sequences are then assigned the correct weights (Table 1). This method is called "modified Voronoi" (mVOR).

Table 1. Comparison of the weighting methods on four simple, contrived examples

Alignment	True	VA	ACL	VOR	mVOR	SS
A	0.500	0.500	0.500	0.501	0.496	0.500
B	0.500	0.500	0.500	0.499	0.504	0.500
A	0.250	0.250	0.250	0.251	0.248	0.290
A	0.250	0.250	0.250	0.251	0.248	0.290
B	0.500	0.500	0.500	0.498	0.504	0.410
AA	0.25	0.25	0.25	0.291	0.25	0.290
AA	0.25	0.25	0.25	0.291	0.25	0.290
BB	0.5	0.5	0.5	0.418	0.50	0.410
AA	0.1667	0.1875	0.1667	0.1842	0.1640	0.1910
AA	0.1667	0.1875	0.1667	0.1842	0.1640	0.1910
BB	0.1667	0.1875	0.1667	0.1854	0.1702	0.1910
BB	0.1667	0.1875	0.1667	0.1854	0.1702	0.1910
CC	0.3333	0.25	0.3333	0.2607	0.3315	0.2361

A half line space separates each alignment. The true weights are based on the criteria given in the text. Due to the Monte Carlo algorithm used to implement the modified "Voronoi" (mVOR) method, the results differ slightly from the true values.

Weights and Phylogeny

The ACL method resembles least-squares estimation of a mean from biased samples (generally a weighted average of sample values). A normal average suffices for independent values with equal variance. Correlated samples are more problematic. Two data points with a high covariance contribute less information concerning the mean than do uncorrelated ones and should receive less weight. Measurements with a high variance are less reliable and should be downweighted. When sample correlation is summarized in a variance–covariance matrix A the weight vector for the weighted average is proportional to $A^{-1}\mathbf{1}$ (24). The ACL scheme (14) follows an idea of Felsenstein (2, 29). Imagine electrical current flows from the root of the tree down the edges and out the leaves. If the edge lengths are proportional to their electrical resistances, current flowing out each leaf equals the leaf weight. Leaves far from the root receive low weights due to greater resistance along that path. When a leaf is duplicated, half as much current flows through each copy. This is the justification for dividing weights when a sequence occurs more than once (and inverting the degenerate matrix is impossible).

For sequences, the ACL method uses a substitute for the variance–covariance matrix. A rooted evolutionary tree is obtained from the given distance matrix. The root is the point of interest: longer branches are less reliable estimators. Branch length from the root to the leaf is the variance of that species. Further, two branches sharing a long common branch (proportional to the covariance) from the root carry much the same information. A species receives lower weight when far from the root or when it has "close neighbors" in the tree. The formalism is exactly the one described for correlated measurements: for a variance–covariance matrix A the weight-vector is a multiple of $A^{-1}\mathbf{1}$. This is formally analogous to the weighting method of the last section that positioned the centroid at equal distances from the sequences.

Recall the definitions of ultrametric and additive trees (17). Briefly, *ultrametric* means rooted with all leaves (=species) equidistant from the root. *Additive trees* may have leaves at different levels, and the root must be found by an independent method. *When a distance matrix* D *allows for exact representation as an ultrametric tree* T, *then the weights calculated via inversion of the variance–covariance matrix* A *of the tree* T *are the same as obtained by inversion of the original matrix* D. To see this, consider how the variance–

covariance matrix A is derived from the tree T. A tree with height h has main diagonal values h. The distance from the root to a branchpoint for sequences i and j will be $h - d_{ij}/2$—i.e., the matrix entry (i, j). Because $d_{ii} = 0$, the matrix A can be written as $A = [h - d_{ij}/2]_{i,j} = (h)_{i,j} - \frac{1}{2}D$. Applying this to weight vector $w = D^{-1}\mathbf{1}$, we obtain a vector all elements of which are $h \cdot \Sigma_i w_i - \frac{1}{2} = h - \frac{1}{2}$; this shows that w is also a correct weight vector for A.

Discussion and Conclusions

To evaluate the methods there are two stances: either that there are no globally objective criteria and a method can be judged only in terms of its efficacy for a particular task or that criteria exist that are self-evident and that should be satisfied. We adopt the second stance because it avoids the tautology inherent in the first and such objective criteria exist. A similar axiomatic approach was applied to clustering methods (26). Define two or more sequences "symmetrical with respect to a distance matrix" if exchanging those sequences does not alter the distance matrix. Symmetrical sequences include, but are not limited to, identical sequences. We believe the following criteria should be met by a weighting method:

C1: Symmetrical sequences should receive equal weights.

C2: If the alignment contains n identical sequences, each with weight x, then removal of all but one should result in the one remaining having weight nx.

C3: Similar sequences should be down-weighted relative to more isolated sequences.

C4: The method should use no unnecessary assumptions and not unnecessarily discard any information ("Occam's razor").

Additional biological information may alter these criteria. Specifically criterion C2 may be relaxed if the number of representatives of a sequence adds to its importance.

Examples

In Table 1 "alignments" 2, 3, and 4 contain duplicate sequences. ACL method cannot be applied directly because the matrix does not invert. Therefore, duplicates were removed before calculation, and resulting weights were divided by the number of duplicates as per criterion C2. SS method oscillates in examples 2 and 3, but modifying the iteration results in the correct eigenvector. For these simple cases there are obvious additive tree representations of the data on which the ACL method was based. In these examples the mVOR and ACL methods conform to the criteria. The SS method results in larger weights for duplicated sequences, a behavior corresponding to a relaxed version of criterion C2.

Fig. 2 (27) illustrates a hazard in requiring a tree for the weighting scheme. The back mutation creates particular problems in tree construction (and back mutations occur in real organisms). Two sets of distances can be assigned to the edges on the tree: ones that reflect the process of evolution and ones calculated from the leaves alone. Tree-independent methods all give similar results, but the ACL method gives results strongly influenced by the method of tree construction (Table 2).

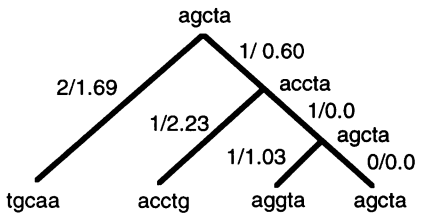

FIG. 2. Model tree redrawn from ref. 27; note the back mutation. Edge distances are the actual number of substitutions (top) and those estimated by Li (27) from the observed sequences (bottom). With one exception, the observed distances between the leaves can be obtained from the tree (e.g., from acctg to tgcaa is four substitutions). In the case of aggta to tgcaa, the observed distance is 3, whereas the actual distance (number of substitutions that actually occurred) is five. This is due to the back mutation. In this case the root is known exactly, and the tree is not ultrametric.

Table 2. Results obtained for the case in Fig. 2

Alignment	VA	VOR	mVOR	SS	ACL actual	ACL calc.
AGCTA	0.1667	0.1321	0.1225	0.1792	0.2857	0.7380
AGGTA	0.2333	0.2791	0.2488	0.2447	0.0	0.0
ACCTG	0.3	0.2995	0.3115	0.2880	0.2857	0.0
TGCAA	0.3	0.2891	0.3171	0.2880	0.4286	0.2620

This is a hypothetical alignment of four 20-nt sequences whose various positions are shown and is taken from Li (27). There is more than one way to produce a tree relating all the sequences. In this table the methods that use only observed distance data provide similar results. The results for the ACL approach use two different trees: (*i*) a tree based on the actual number of substitutions that occurred (Fig. 2), information not normally accessible to the researcher, and (*ii*) a tree based on distances as calculated (calc.) by using the method of Li (ref. 27, see figure 2 of that paper).

If the data fit a tree well, the ACL method gives results similar to the other methods. This situation is illustrated for 10 5S RNA sequences, a tree for which is shown in Fig. 3. Sequence data, alignment, the tree, and reconstructed ancestral sequences assigned to the internal nodes are taken from ref. 28. We used the internal sequences to construct an additive distance matrix and derived a variance–covariance matrix for the ACL method. We also constructed an ultrametric tree approximating the distances between the sequences by using the program KITSCH (29). The tree (Fig. 3) topology given by ref. 28 matches that from KITSCH. Branch lengths are shown for the ultrametric and additive tree, but only the ultrametric tree is roughly to scale. There is good agreement between the methods (Table 3). ACL (ultrametric tree) gives equal weights to adjacent species (e.g., 9 and 10). In the additive tree 10 is farther away than 9 from their common ancestor, and 10 receives a lower weight (ACL). The distance-based methods all perceive 10 to be more important than 9. This is due to different philosophies.

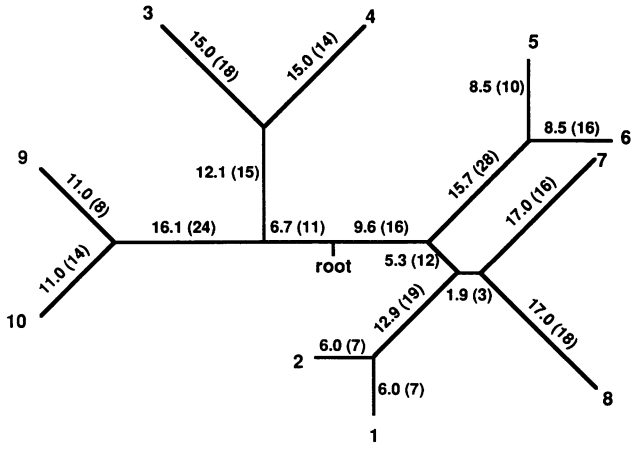

FIG. 3. A tree based on an alignment of 10 5S RNA sequences (28). Distances on the edges of the tree are ultrametric (first number) and additive (second number) approximations, respectively. See text for details. The numbers at the leaf positions correspond to organisms as follows: 1, *Auricularia auricula-judae*; 2, *Auricularia edulis*; 3, *Bacillus brevis*; 4, *Bacillus firmus*; 5, *Equisitum arvense*; 6, *Cycad revoluta*; 7, *Caenorhabditis elegans*; 8, *Gallus gallus*; 9, *Jungermannia subulata* chloroplast; 10, *Dryopteris acuminata* chloroplast.

Table 3. Results obtained for the 5S RNA alignment described in Fig. 3

Sequence	VA	VOR	mVOR	SS	ACL ultra	ACL additive
1	0.0962	0.0840	0.0900	0.0977	0.0627	0.0575
2	0.0925	0.0763	0.0798	0.0942	0.0627	0.0411
3	0.1061	0.1155	0.1142	0.1045	0.1307	0.1330
4	0.1007	0.1019	0.1033	0.0997	0.1307	0.1710
5	0.0958	0.0932	0.0915	0.0968	0.0919	0.0850
6	0.0977	0.0980	0.0974	0.0988	0.0919	0.0850
7	0.0914	0.0864	0.0917	0.0929	0.0958	0.0998
8	0.0934	0.0999	0.0959	0.0950	0.0958	0.0888
9	0.1106	0.1121	0.1123	0.1076	0.1186	0.1520
10	0.1156	0.1328	0.1238	0.1122	0.1186	0.0870

Sequence numbers are the same as in Fig. 3.

The methods based solely on the (dis)similarity matrix between the sequences fulfill criterion C1. mVOR fulfills the weaker criterion of assigning equal weights to copies of identical sequences. Current data suggest that mVOR also fulfills C1 for symmetric sequences. ACL and mVOR obey criterion C2. SS method tends to give slightly higher weights to duplicated sequences than C2 would suggest (possibly desirable). Methods VA and VOR are inconsistent regarding C2 (Table 1). Criterion C3 appears satisfied by all the methods. Regarding criterion C4, there are considerable differences. The examples show that the assumption that the data can be related by a tree and that the researcher can produce an appropriate tree can be risky. However, a tree facilitates addition of accessory information.

One advantage of the ACL method (14) is a particular point in a tree can be chosen as root, and leaves can be weighted relative to that point. One can weight a point in a subtree of rodents within a larger tree of mammals while taking into account the mammal sequences. A similar effect can be obtained with the distance methods by using only the rodent sequences, but the mammal information is not exploited. This effect violates criterion C4: it discards potentially useful information, but it is unclear how significant this is.

Of the methods not requiring a phylogeny, the mVOR method scores best. It is more often correct than the other distance methods of Table 1. A shortcoming is that the algorithm is stochastic. An analytic formulation of the Voronoi weights is an open problem. The mVOR method is especially appropriate where convergence is a reasonable hypothesis, the phylogeny is dubious, or the data fit a tree poorly. The VOR method weights sequences such that diversity is estimated. Rare outlying sequences are upweighted to represent a large part of the space "on their own." Altschul *et al.* (14) down-weight sequences far from the root because they convey less information about the root. The ACL method answers the question "How should the sequences be weighted to best estimate a particular point?" The VOR and mVOR methods answer the question "How should the sequences be weighted to best estimate the diversity of the group of sequences?" Therefore, the mVOR method and ACL method are the best available, although for different problems.

We thank Andreas Dress for helpful comments. M.V. was supported by National Science Foundation Grants DMS 90-05833 and DMS 87-20208 and National Institutes of Health Grant GM-36230.

1. Sneath, P. H. A. & Sokal, R. R. (1973) *Numerical Taxonomy* (Freeman, San Francisco).
2. Felsenstein, J. (1973) *Am. J. Hum. Gen.* **25**, 471–492.
3. Hogeweg, P. & Hesper, B. (1984) *J. Mol. Evol.* **20**, 175–186.
4. Waterman, M. S. & Perlwitz, M. D. (1986) *Bull. Math. Biol.* **46**, 567–577.
5. Lipman, D. J., Altschul, S. F. & Kececioglu, J. D. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 4412–4415.
6. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 4355–4358.
7. Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987) *J. Mol. Biol.* **195**, 957–961.
8. Barton, G. J. & Sternberg, M. J. E. (1987) *J. Mol. Biol.* **198**, 327–337.
9. Higgins, D. G. & Sharp, P. M. (1989) *Comput. Appl. Biosci.* **5**, 151–153.
10. Vingron, M. & Argos, P. (1989) *Comput. Appl. Biosci.* **5**, 115–121.
11. Taylor, W. R. (1988) *J. Mol. Evol.* **28**, 161–169.
12. Corpet, F. (1988) *Nucleic Acids Res.* **16**, 10881–10890.
13. Sibbald, P. R. & Argos, P. (1990) *J. Mol. Biol.* **216**, 813–818.
14. Altschul, S. F., Carroll, R. J. & Lipman, D. J. (1989) *J. Mol. Biol.* **207**, 647–653.
15. Sander, C. & Schneider, R. (1991) *Proteins Struct. Funct. Genet.* **9**, 56–68.
16. Dress, A. & von Haeseler, A., eds. (1990) *Trees and Hierarchical Structures*, Lecture Notes in Biomathematics (Springer, Heidelberg).
17. Barthélemy, J.-P. & Guénoche, A. (1991) *Trees and Proximity Representations* (Wiley, New York).
18. Eigen, M., Winkler-Oswatitsch, R. & Dress, A. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 5913–5917.
19. van Heel, M. (1991) *J. Mol. Biol.* **220**, 877–887.
20. Higgins, D. G. (1992) *Comput. Appl. Biosci.* **8**, 15–22.
21. Gantmacher, F. R. (1959) *The Theory of Matrices* (Chelsea, New York).
22. Stoer, J. & Bulirsch, R. (1991) *Introduction to Numerical Analysis* (Springer, Heidelberg).
23. Bickel, P. J. & Doksum, K. A. (1977) *Mathematical Statistics* (Holden–Day, New York).
24. Strang, G. (1986) *Introduction to Applied Mathematics* (Wellesley–Cambridge, Cambridge, MA).
25. Altschul, S. F. & Lipman, D. J. (1990) *Nature (London)* **348**, 493–494.
26. Bandelt, H.-J. & Dress, A. W. M. (1989) *Bull. Math. Biol.* **51**, 133–166.
27. Li, W.-H. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 1085–1089.
28. Hein, J. (1990) *Methods Enzymol.* **183**, 626–645.
29. Felsenstein, J. (1991) *Cladistics* **5**, 164–166.