Comparative biosequence metrics

Article *in* Journal of Molecular Evolution · February 1981

DOI: 10.1007/BF01733210 · Source: PubMed

3 authors, including:

Temple F. Smith
Boston University
259 PUBLICATIONS **23,438** CITATIONS

SEE PROFILE

J Mol Evol (1981) 18:38–46

Journal of Molecular Evolution

Comparative Biosequence Metrics

T.F. Smith[1], M.S. Waterman[2], and W.M. Fitch[3]

1 Northern Michigan University, Marquette, Michigan 49855, USA
2 Los Alamos Scientific Laboratory, Los Alamos, New Mexico 87545, USA
3 University of Wisconsin, Madison, Wisconsin 53706, USA

Summary. The sequence alignment algorithms of Needleman and Wunsch (1970) and Sellers (1974) are compared. Although the former maximizes similarity and the latter minimizes differences, the two procedures are proven to be equivalent. The equivalence relations necessary for each procedure to give the same result are: 1, the weight assigned to gaps in the Sellers algorithm exceed that in the Needleman-Wunsch algorithm by exactly half the length of the gap times the maximum match value; and 2, for any pair of aligned elements, the degree of similarity assigned by the Needleman-Wunsch algorithm plus the degree of dissimilarity assigned by the Sellers algorithm equal a constant. The utility of the algorithms is independent of the nature of the elements in the sequence and could include anything from geological sequences to the amino acid sequences of proteins. Examples are provided using known nucleotide sequences, one of which shows two sequences to be analogous rather than homologous.

Key words: Analogy – Convergence – Distance – Divergence – Homology – Needleman-Wunsch-algorithm – Sellers-algorithm – Sequence-alignment

Introduction

Currently there are two major algorithms in the literature directly applicable to comparing unaligned macromolecular sequences. These are the algorithms of Needleman and Wunsch (1970) and of Sellers (1974), the latter as generalized by Waterman, Smith and Beyer (1976). Both are designed to produce an optimum measure between any two sequences as a function of the minimum number of changes required to convert one into the other.

Both may be viewed as an extension of the original Hamming (1950) sequence metric idea. The extension includes deletions and insertions as allowed changes in sequence elements, in addition to a change in the character state of an element. Still excluded from analysis are duplication and inversion events. Fortunately, these latter two types of events, while well known for intergene or inter-chromosomal events, are rare for the intragene mutational histories so far investigated (see Fitch 1977 for a documented exception; intervening sequences within genes may also prove to be an exception).

There are two major differences between the Needleman-Wunsch and the Sellers algorithms. The most obvious is that the Needleman-Wunsch algorithm results in alignments having a maximum similarity measure, while the Sellers algorithm results in alignments having a minimal distance or metric measure of dissimilarity. The second major difference between them is in their origin. The first was the result of a heuristic approach to an important biological problem, while the second was the result of a search for a rigorous mathematical solution for the problem.

It is the main purpose of the present study to demonstrate the distinctive as well as similar characteristics of these two methods for comparative analysis and, in particular, to note the existence of a set of conditions resulting in their equivalence. The conditions for equivalence are of great importance in understanding the applicability of these tools to various problems currently under investigation in molecular biology.

The Measures

In order to facilitate the comparisons between the two algorithms we will describe them rigorously using com-

Offprint requests to: W.M. Fitch, Department of Physiological Chemistry, 1215 Linden Drive, 589 Medical Sciences Building, Madison, WI 53706, USA

0022–2844/81/0018/0038/$ 01.80

patible notation. The actual matrix procedures used to implement these algorithms on modern computers is given in the next section.

Consider two sequences, nucleotide sequences for simplicity, A and B of length n and m respectively, where $\underset{\sim}{A} = a_1\ a_2\ a_3\ \ldots\ a_n$ and $\underset{\sim}{B} = b_1\ b_2\ b_3\ \ldots\ b_m$. An alignment between two such sequences is defined as an ordered sequence of pairs, each pair containing one element from each sequence or an element from either sequence and the null element, with the order of the original sequences preserved. Deletions or insertions (gaps)[1] are indicated by alignment pairs containing a null element Δ. For example, the alignment:

$$\begin{array}{l} C\ A\ U\ G\ \Delta\ \Delta\ \Delta\ C\ A\ \Delta\ \Delta \\ \underline{C}\ C\ \underline{U}\ \underline{G}\ U\ C\ C\ \underline{C}\ \underline{A}\ U\ U \end{array} \qquad \text{(A1)}$$

between two short nucleotide sequences contains two gaps, one of length three and one of length two. It also has five matching elements shown by underlining.

Now for a given alignment, let λ be the number of pairs containing identical or matched elements, μ be the number of pairs containing non-identical or mismatched elements (excluding the null element) and Δ_k be the number of gaps of length k. Each match and each mismatch involves two elements one from each sequence, while each gap of length k involves only k elements from one sequence. Now since there are a total of n plus m elements in the two sequences,

$$n + m = 2(\lambda + \mu) + \sum_k k\,\Delta_k \qquad (1)$$

It must be noted that any unmatched terminal subsequence is always associated with a gap in the above definition of an alignment. This was not explicitly assumed in the original Needleman-Wunsch algorithm, but will initially be assumed here for simplicity. We will return to the less restrictive case later. Under this restriction the Needleman-Wunsch algorithm results in the alignment(s) given by *a similarity measure, s,* where

$$s = \text{maximum}\left[\lambda - \sum_k w_k\,\Delta_k\right] . \qquad (2)$$

where w_k is the weight or "penalty" associated with a gap of length k and the maximum is over all possible alignments. The quantity, s, is thus a measure of maximum similarity minus gap penalties.

The Sellers algorithm on the other hand results in the alignment(s) given by *a distance measure, d,* where

$$d = \text{minimum}\left[\mu + \sum_k w'_k\,\Delta_k\right] . \qquad (3)$$

Here w'_k is a gap weight analogous to w_k and d is a proper distance obeying the required metric properties. These are: for any two sequences A and B:

$0 \leqslant d(A, B)$ for all sequences A and B and zero if and only if sequence $\underset{\sim}{A}$ is identical to $\underset{\sim}{B}$;

$d(\underset{\sim}{A}, \underset{\sim}{B}) = d(\underset{\sim}{B}, \underset{\sim}{A})$ for all sequences $\underset{\sim}{A}$ and $\underset{\sim}{B}$;

$d(\underset{\sim}{A}, \underset{\sim}{B}) \leqslant d(\underset{\sim}{A}, \underset{\sim}{C}) + d(\underset{\sim}{C}, \underset{\sim}{B})$ for any third sequence $\underset{\sim}{C}$.

The last relation, the triangle inequality, appears critical if comparative sequence distances are to be used for taxonomic reconstructions.

An equivalence between the Needleman-Wunsch and the Sellers algorithms is established via equation (1). Using equation (1) the maximization given in (2) can be rewritten as

$$s = \text{maximum}\left\{\frac{n+m}{2} - \mu - \frac{1}{2}\sum (k + 2w_k)\,\Delta_k\right\}. \qquad (4a)$$

The alignment independent term may be moved outside the maximum to give

$$s = \frac{n+m}{2} + \text{maximum}\left\{-\mu - \sum (k/2 + w_k)\,\Delta_k\right\}. \qquad (4b)$$

Next, the maximum of a quantity can be replaced by the negative of the minimum of its negative to give

$$s = \frac{n+m}{2} - \text{minimum}\left\{\mu + \sum (k/2 + w_k)\,\Delta_k\right\}. \qquad (4c)$$

Thus the alignments obtained for the Needleman-Wunsch maximum similarity, s, are identical to those obtained from the minimization of the quantity

$$\mu + \sum_k (k/2 + w_k)\,\Delta_k .$$

This is identical to the quantity minimized by the Sellers algorithm if the Sellers' gap weight w'_k is equated to the Needleman-Wunsch gap weight plus half the gap length[2].

$$w'_k = k/2 + w_k \qquad (5)$$

If the gap weight for the similarity measure were a constant independent of gap length, the gap weight for the equivalent distance measure would contain a term equal to half the gap length.

The logic leading to relationship (5) can be generalized to include matches and mismatches of varying degree[3]. Equation 1 can be written as

1 While the term deletion might be mathematically preferable to describe a subsequence of k null elements, it is not determinable from a single pair of sequences whether the gap was produced by the insertion of k elements in one sequence or the deletion of k elements from the other. Thus we shall employ the indifferent term, gap

2 Equation (5) would require, for some values of w'_k, the equivalent w_k to become negative for large k. This, in fact, sets a lower bound on w'_k as a function of k, namely, $w'_k \geqslant k/2$ for all k

$$n + m = 2 \sum_i \eta_i + \sum_k k\Delta_k$$

where η_i is the number of matches of degree i in an alignment. Here we consider each pair of associated elements to "match", but to various degrees, α_i. Letting a perfect match have a maximum similarity of α_{max} and others a lesser degree down to a complete mismatch having a similarity of zero, one has $0 \leqslant \alpha_i \leqslant \alpha_{max}$. The Needleman-Wunsch similarity measure can still be shown equivalent to the Sellers metric. If, in equation 2, we assign a weight α_i to each match of degree i, then the similarity measure becomes

$$s = \max \{\Sigma \alpha_i \eta_i - \Sigma w_k \Delta_k\} . \quad (6a)$$

In a similar manner a weight or distance β_j can be assigned measuring the degree of mismatch in the Sellers distance for any degree of match.

$$d = \min \{\Sigma \beta_j \eta_j + \Sigma w'_k \Delta_k\} . \quad (6b)$$

Here β_j is just a distance measure between paired elements in the alignment. Letting the relationship between α_i and β_i be

$$\alpha_i = \alpha_{max} - \beta_i \ , \quad (7)$$

will yield an equivalence between the optimal alignments produced by the two algorithms provided that the gap weights are related, as before, by

$$w'_k = w_k + k\,\alpha_{max}/2$$

Consider the case where the sequence lengths are equal (m = n) and $w_k \geqslant 0$ (hence $w'_k \geqslant k\alpha_{max}/2$). The minimum and maximum values of s are 0 and $n\alpha_{max}$. The corresponding values of d are 0 and $n\beta_{max}$. If the weights have been chosen to make the algorithms equivalent, they have the same bounds and for any given alignment, $s + d = n\alpha_{max}$. If the sequences are not of the same length, the lower bound of s is $-(m-n)w_{m-n}$ where m is the larger of the two lengths. This can be seen in an example of aligning poly A_m to poly T_n where there must be n A-T mismatches and m-n A's paired with the null element. The same phenomomon will raise the upper bound of d by the same amount to $n\beta_{max} + (m-n)w'_{m-n}$. Thus, if the weights are equivalent, s + d remains equal to $n\alpha_{max}$.

The Algorithms

Both the Sellers metric and the Needleman-Wunsch similarity measure have simple iterative algorithms, particularly easy to adapt to computer analysis. As in the last section, both will initially be presented for the simplest case in which any unmatched terminal subsequence is associated with a gap.

The calculation of the Needleman-Wunsch similarity measure between two sequences A and B of length n and m require an n + 1 by m + 1 matrix, S. The following equation gives the algorithm for calculating element S_{ij}

$$S_{ij} = \max \{S_{i-1,j-1} + \alpha(a_i, b_j); \max_{k \geqslant 1} [S_{i,j-k} - w_k]; \max_{k \geqslant 1} [S_{i-k,j} - w_k]\} . \quad (8)$$

Here $\alpha(a_i, b_j)$ is one of α_i values reflecting the degree to which the ith element of sequence A matches the jth element of B. The algorithm is initialized by setting $S_{0k} = S_{k0} = -w_k$. Equation (8) follows from the fact that in any alignment, a_i is either associated with b_j,

$$S_{i-1,j-1} + \alpha(a_i, b_j) \ ,$$

or a null element in a gap;

$$S_{i,j-k} - w_k$$

with weight w_k, and finally element b_j is either associated with a_i or a null element in a gap,

$$S_{i-k,j} - w_k \ .$$

The maximum similarity measure, with weighted unmatched end sequences is just the value in the last calculated element $S_{n,m}$.

The original Needleman-Wunsch measure did not weight any unpaired terminal subsequences, in other words they were not associated with a terminal gap in the other sequence. This original measure is obtainable from equation (8) if the matrix is initialized to $S_{0k} = S_{k0} = 0$ rather than to $-w_k$. Under these conditions the measure is the maximum valued matrix border element from $S_{m,}$ and $S_{,n}$. The relationship of this measure to the Sellers metric will be discussed below.

In a manner analogous to the definition of the similarity matrix S, Sellers defined a distance matrix D. The algorithm for its generation is given by

$$D_{ij} = \min \{D_{i-1,j-1} + \beta(a_i, b_j) \ , \min_k [D_{i,j-k} + w'_k], \min_k [D_{i-k,j} + w'_k]\} . \quad (9)$$

3 In protein sequences for example, the amino acids glycine and alanine could be considered to have a match value of 2 or a mismatch value of 1 based on the genetic code. Any basis (be it chemical or structural) could be used to generate values. Integral values are not required. Recognizing that transition mutations (A↔G, C↔U) are about twice as common as transversions relative to their expected frequencies, one could consider AG or CU mismatches at a value of 0.75 while all others were set at 1.5 so that the sum of the mismatches has the same expectation

The algorithm is initialized by setting $D_{0k} = D_{k0} = w_k'$. Equation (9) follows from logic identical to that for equation (8). And as in the similarity measure algorithm, the Sellers metric distance is given by the last calculated element D_{nm}.

Sellers (1979) noted, as in the original Needleman-Wunsch case, that the distance matrix can be initialized with all zeroes or zeroes associated only with the longer of the two sequences under analysis. This latter suggestion is important, for the proper question may not be to find the minimum distance between a short fragment and a long sequence, but rather to find a segment of the longer sequence which is minimum distance from the fragment. This, Sellers (1979) showed, was accomplished by initializing $D_{0\ell} = 0$ for all ℓ associated with the longer sequence and $D_{k0} = w_k'$ for k associated with the shorter fragment. The distance is then given by the minimum $D_{n\ell}$ value[4].

The alignments associated with any of the above measures are obtained by tracing back through the S (or D) matrix to obtain the correlated pairs. Algorithmically, this can be done simply by calculating a second matrix simultaneously with the first. The elements of the second matrix record which terms in equations (8) (or (9)) contributed to the calculation of S_{ij} (or D_{ij}). In particular, if this is done on a computer with "packed words" the length of the involved deletion(s) can be recorded as well.

Before leaving this discussion of the algorithms, it is important to note the mathematical consequences of not weighting unpaired terminal subsequences. Suppose three short sequences of equal length, $\underset{\sim}{A}$, $\underset{\sim}{B}$, and $\underset{\sim}{C}$, have optimal alignments associated with Sellers distances D(A,B), D(A,C) and D(B,C) that require no deletions. Consider two longer sequences $\underset{\sim}{A}'$ and $\underset{\sim}{B}'$ containing as end sequences the $\underset{\sim}{A}$ and $\underset{\sim}{B}$ sequences respectively. Now given the triangle inequality holds for $\underset{\sim}{A}$, $\underset{\sim}{B}$, and $\underset{\sim}{C}$, and the unpaired terminal sequences among $\underset{\sim}{A}'$, $\underset{\sim}{B}'$, and $\underset{\sim}{C}'$ go unweighted, one could have

$$D(A', C) = D(A, C),$$
$$D(B', C) = D(B, C),$$

and

$$D(A', B') > D(A, B),$$

which permits

$$D(A', B') > D(A', C) + D(B', C).$$

This possible violation of the triangle inequality raises some question as to the mathematical interpretation of measures that do not weight unmatched terminal subsequences.

The dependence of the optimal alignment on the value assigned to w_k (and hence to w'_k) is illustrated in the following example. Consider the sequences GCAGAGCACU and GCUGGAAGGCAU. If unpaired terminal subsequences are weighted (associated with null elements), then, for all $w_k \geqslant 1.0$, the alignment

$$\begin{array}{cccccccccccc} G & C & A & G & \Delta & \Delta & A & G & C & A & C & U \\ \underline{G} & \underline{C} & U & \underline{G} & G & A & \underline{A} & \underline{G} & G & C & A & \underline{U} \end{array} \qquad \text{(A2)}$$

is obtained. This alignment contains six matches and an internal gap of length two. The traceback giving this alignment is identified in Fig. 1 which displays the equation 8, D matrix for this example. If unpaired terminal subsequences are not weighted, not associated with a gap, then the alignment,

$$\begin{array}{cccccccccccc} & & G & C & A & G & A & G & C & A & C & U \\ G & C & U & \underline{G} & G & \underline{A} & A & G & \underline{G} & \underline{C} & \underline{A} & U \end{array} \quad , \qquad \text{(A3)}$$

	Δ	G	C	A	G	A	G	C	A	C	U
Δ	0	1.6	2.2	2.8	3.4	4.0	4.6	5.2	5.8	6.4	7.0
G	1.6	0	1.6	2.2	2.8	3.4	4.0	4.6	5.2	5.8	6.4
C	2.2	1.6	0	1.6	2.2	2.8	3.4	4.0	4.6	5.2	5.8
U	2.8	2.2	1.6	1.0	2.6	3.2	3.8	4.4	5.0	5.6	5.2
G	3.4	2.8	2.2	2.6	1.0	2.6	3.2	3.8	4.4	5.0	5.6
G	4.0	3.4	2.8	3.2	2.6	2.0	2.6	4.2	4.8	5.4	6.0
A	4.6	4.0	3.4	2.8	3.2	2.6	3.0	3.6	4.2	5.8	6.4
A	5.2	4.6	4.0	3.4	3.8	3.2	3.6	4.0	3.6	5.2	5.8
G	5.8	5.2	4.6	5.0	3.4	4.8	3.2	4.6	5.0	4.6	6.2
G	6.4	5.8	5.2	5.6	5.0	4.4	4.8	4.2	5.6	6.0	5.6
C	7.0	6.4	5.8	6.2	5.6	6.0	5.4	4.8	5.2	5.6	7.0
A	7.6	7.0	6.4	5.8	6.2	5.6	6.0	6.4	4.8	6.2	6.6
U	8.2	7.6	7.0	7.4	6.8	7.2	6.6	7.0	6.4	5.8	6.2

Fig. 1. An example D matrix for sequences GCAGAGCACU and GCUGGAAGGCAU showing traceback for optimal alignment A2. The values for β and w_k' used here were, $\beta = 1.0$ if $a_i \neq b_j$ and zero otherwise and $w_k' = 1.0 + 0.6\ k$. Since we are weighting gaps at the ends of the sequence, the traceback begins in the lower right-hand cell which contains the distance measure (6.2). The traceback follows the course shown by the arrows. Although an algorithm is easily written that would find this path from these values, that is computationally very inefficient compared to storing, as the D matrix is created, the cell(s) whose D_{ij} value contributes to the value of an ensuing D value

4 Setting both boundaries to zeroes and identifying the distance measure with the last calculated element D_{nm}, gives a non-weighting to unmatched terminal subsequences only at the beginning of the sequences

which contains only five matching element pairs is obtained for all $w_k > 1.0$.

If the gap w_k is equal to (or just less than) unity, alignment A2 is obtained regardless of whether unmatched terminal subsequences are weighted. If w_k is much less than unity for $k \leqslant 2$, the optimal alignment has eight matches

$$\begin{array}{l} G\ C\ \Delta\ A\ G\ A\ \Delta\ \Delta\ G\ C\ A\ C\ U \\ \underline{G}\ \underline{C}\ U\ G\ \underline{G}\ \underline{A}\ A\ G\ \underline{G}\ \underline{C}\ \underline{A}\ \Delta\ \underline{U} \end{array} \qquad \text{(A4)}$$

This latter alignment represents the degenerate case where gaps have been inserted "at will" to maximize matches. This procedure, which has been used in the past, is of questionable value if genetic history is being investigated. Its most glaring fault is the frequent introduction of two gaps to attain only one extra match.

Finally, before looking at any particular biological applications, the statistical behavior of these measures should be noted. To begin with, there is no known analytical means of calculating the probability of a given number of matches produced by these measures as a function of composition and w_k. Only between infinitely long sequences in which gaps are not considered can we state the expected matching probabilities as simple functions of sequence element composition. Thus Monte Carlo methods seem the best. For example, the distribution of matches among twenty random comparisons between two sequences of length twelve of uniform nucleotide composition calculated for two different gap weightings are given in Table 1.

Table 1. Distribution of Sellers metric as a function of gap weight

A		B	
$w_k' = 0.9 + 0.5\ k$		$w_k' = 1.0 + 0.6\ k$	
freq	distance	freq	distance
1	8.8	1	10.6
1	8.0	1	9.8
4	7.8	1	9.4
3	7.7	1	9.2
2	7.6	3	9.0
2	7.0	3	8.4
3	6.8	1	8.2
1	6.7	3	8.0
1	6.6	1	7.6
2	6.0	1	7.2
		3	7.0
		1	6.4
mean±std. div.	7.3±0.8		8.3±1.1

These distances were calculated by equation (9) for the two weighting functions given, between the sequence ATCGATCGATCG and twenty random shufflings of the same. The average distance increases with w_k' as expected. The large deviations here are in part a result of the shortness of the sequences used, but as noted in the text, the sigmas for longer sequences, up to one hundred, still have values near ten percent. Equivalent w_k values are 0.9 and 1.0 + 0.1 k for A and B respectively

Applications

Assigning gap weights and deciding whether or not to weight unpaired terminal subsequences are important considerations in the application of these algorithms. There are two general rules. First, if one is attempting to obtain a measure of genetic distance between two equivalently defined sequences (presumably homologous), all unpaired elements should be weighted as gaps.

The second general rule arises from the following considerations. The evolutionarily effective substitution of one nucleotide for another is considerably more common than the successful insertion or deletion of one or more nucleotides. Thus the "penalty" for a gap should be greater than the value of an added match achieved or a single mismatch avoided. The same rationale also requires, for the same degree of matching, that a single gap of k residues should be preferred to multiple gaps with a total of k null elements. The second general rule then is to make $w_k > \alpha_{max}$. However, beyond these two rules there are no obvious fixed and hard rules for assigning these weights.

A good example of an analysis requiring additional nonsequence information to resolve these problems is found in the recent work on the cytochrome's c by Dickerson (1980a, 1980b). Direct application of the Sellers or Needleman-Wunsch algorithms to the broad range of taxonomic units analyzed results in little consistency and low homology. This is, perhaps, to be expected considering the time since divergence of the various lines of descent. However, if the various structurally equivalent sites within these functionally equivalent molecules are identified and the sequence comparisons constrained within them, considerable interpretable homology is observed. For example, there is a cysteine-histidine amino acid pair in the first third of all cytochromes and an associated methionine in the last third. These sites are absolutely functionally equivalent as the heme-binding side-chain amino acids.

The incorporation of such functionally prealigned sequence elements is rather straightforward. For example, the cytochrome sequences studied by Dickerson can be expressed as

$$X_1 \ldots X_{n-1}\ X_n\ CH\ Y_1 Y_2 \ldots Y_{m-1}\ Y_m\ M\ Z_1 Z_2 \ldots Z_k$$

three subsequences separated by the cysteine-histidine pair and the methionine, heme binding amino acids. A sequence comparison between any pair of cytochromes, C and C′, thus reduces to three shorter sequence comparisons: a "Y" or mid sequence comparison in which the D or S matrix boundaries are set to weight unpaired terminal subsequences as deletions

```
5'  Δ Δ T C T C Δ Δ[27 NUCLEOTIDES]G A A C A C G C C G C C T C T T C T  3'
5'  G C G G T C A G[NO NUCLEOTIDES]Δ Δ Δ T C C G C C G A C T C T Δ Δ Δ  3'
```
(A5)

```
5'  T C T C[27 NUCLEOTIDES]G A A C A C G C C G C C T C T T C T  3'
5'  G C Δ Δ[NO NUCLEOTIDES]G G T C A G T C C G C C G A C T C T  3'
```
(A6)

or gaps, the value of the measure is given by $D_{m,m'}$ or $S_{m,m'}$; a "X" or leading subsequence comparison in which the D or S matrix boundaries are set to zeroes but where the measure is still given by the $D_{n,n'}$ or $S_{n,n'}$ matrix element (this has the effect of not weighting as gaps unpaired initial elements but weighting any trailing gaps between the "X" region, and the fixed CH amino acid pair.); finally a "Z" terminal subsequence comparison in which the D or S matrix is initialized to weight all initial unpaired elements as gaps but where the measure is the minimum (as a function of ℓ) $D_{k,\ell}$ or $D_{\ell,k}$ matrix element, thus not weighting the trailing unpaired elements.

The cytochrome-c studies of Dickerson also demonstrate a second problem in the choice of algorithm parameters. This is in choosing the distance measure β between sequence elements and its corresponding similarity α given in equation (7). Initially one is tempted to measure the distance between the amino acids in terms of the underlying genetic code. However, with the additional information on the molecule's three dimensional structure and the relative high probability (over the taxonomic range from blue green algae to tuna) of multiple nucleotide substitutions at any given site in the cytochrome gene, a measure of amino acid distance in terms of their physical-chemical similarity proved rather useful (Dickerson 1980b). This is no doubt a result of the fact that the structural requirements of the various protein regions constrains, to some degree, the acceptable amino acid replacements. One might therefore well use a measure of distance obtained by subtracting the values of the log-odds matrix (Fig. 84, Dayhoff et al. 1978) from 17 to get a set of values ranging from zero to 25 that are related (exponentially) to the likelihood of finding a particular pairing of amino acids, if they indeed had had a common ancestor.

In a recent comparison of phage DNA replication initiation regions (Sims et al. 1979), an alignment was obtained by keying three nucleotide positions. These were the actual replication origin, a 5′ adjacent gene terminator and a 3′ adjacent AUG initiation codon. Unfortunately, the alignment procedure used was stated as "inserting occasional gaps to maximize homology[5]."

5 The word homology here refers only to matches and therefore is in a strict sense not used as a measure of total similarity as measured by the S matrix given by equation (8)

Such a procedure is capable in the extreme of maximizing the number matches. However, even this goal can rarely be optimized by hand. For example, the alignment, A5, proposed by Sims et al. (1979, Fig. 9) for a region of the H gene of st-1 (upper row) and G4 (lower row) phase, contains 4 gaps and 11 matches. Using the Sellers D metric and not weighting as gaps unpaired terminal subsequences at the 3′ (rightward) end one obtains the alignment, A6, which contains 12 matches and only one internal gap which was weighted rather heavily at w'_k equal to 1.00 + 1.10 k. Thus the simple "looks reasonable" approach is clearly unacceptable.

Finally there is the recent study by Rosenberg and Court (1979) aligning 46 antisense strand promotor regions with no gaps or deletions. Here again an attempt was made to key the alignment to an invariant position, the "invariant T" in the sixth position of the so called "Pribnow box", TATAATG (Pribnow 1975). While these alignments are approximately correlated with a function site, that of mRNA initiation, the presumed start site is not always clearly known and there was permitted to be anywhere from four to eight nucleotides between the invariant T and the first nucleotide transcribed. A number of the proposed alignments show only two additional matches in the Pribnow heptamer given the 'invariant T', although over the entire data set analyzed by Rosenberg and Court, the alignment on this position leads to a high overall correlation with other Pribnow positions.

Such statistics do not imply directly any significance or lack thereof to the "Pribnow box" but do point out difficulties of keying an alignment on a site of uncertain functional equivalence. A number of alignments obtainable from among the known 46 promotor regions studied by Rosenberg and Court (1979) demonstrate the problems of alignments without reference to well fixed 'key' positions. For example, the λ PRE region (Rosenberg et al. 1978) contains three potential Pribnow sequences, each with four matches (see alignment A7). None of these "Pribnow" sequences with underlined matches corresponds to the proposed (Rosenberg and Court 1979) subsequence which is overlined, and contains only three matches. This overlined subsequence is preferred by Rosenberg and Court by virtue of its position relative to the starred mRNA transcription initiation site(s), $\overset{**}{A}G$, which is not unambiguously defined. Moreover, even if the initiation site is unambiguously defined, its placement relative to the "in-

```
                     ---------       **
. . . . .AACCATATGTAAGTATTTCCTTAGATAACAATTG . . .
                   TATAATG

                                   TATAATG
                                                          (A7)
                                 TATAATG

                 TATAATG
```

```
                   ---------      *
. . . . . TTTGTTATGCTATGGTTAT. . . . .
                 TATAATG                                  (A8)
               TATAATG
```

```
ATTCTCTTGTTGACATTTTAAAAGAGCGTGGΔATTACTATCTGAGTΔΔΔCCGATΔGCTGTTC        (A9)
     TTAAATAGCTTGCAAAATAΔCGTGGCCTTATGGTTACAGTAGTCCCATCGCAGTTCGC
```

```
                ACACTTTATGCTTCCGGCTCGTATGTTGTGTGGTATTGTGAGCGGATAACAATTTCA
                                                                          (A10)
ACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGG
```

```
                                              ---------      **
ACACTTTATGCTTCCGGCTCGTATGTTGTGTGGTATTGTGAGCGGATAACAATTTCA              (A11)

ACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGG
```

variant T" is not. In the case of Ga1P$_2$, Rosenberg and Court propose that the Pribnow box is as shown in alignment A8 by the overlining. But there is a second Pribnow heptamer shown below the galP$_2$ sequence having one additional matched nucleotide, including the "invariant T". The only drawback to this alternative would appear to be that there are 9 nucleotides between the "invariant T" and the transcription initiation site (shown by the * in A8) while all the other 45 alignments have at most 8 nucleotides between them. We know of no reason, however, that the range of these two sites must be limited to that range and Rosenberg and Court do not claim that that limited range was an a priori assumption in their alignment procedure.

The total promotor regions can be examined for optimal alignments without requiring that there be a T five to nine nucleotides 5′ of a presumed initiation site. For example, ϕX174D and ϕX174B promotor regions give an optimum alignment, A9, with a w_k of 1.0 + 0.6 k and the unpaired ends not weighted. This gives 32 matches, a Sellers' d of 29.6 and a Needleman-Wunsch measure s of 24.4. Neither of these values are more than 0.8 standard deviations away from the mean of measures between the ϕX174D and twenty random shufflings of ϕX174B. Note both the proposed Pribnow sequences have a T associated with the "invariant T". The proposed alignment of Rosenberg and Court, which does not contain any deletions, has 15 matches, almost exactly what is expected for the random case, for the given base composition. This in fact is the rule among both the proposed alignments as well as among three hundred optimal alignments obtained using equation 8. In fact much higher homologies were found among many of the random shuffled sequences!

There are exceptions within this data set of 46 promotor regions to this near pairwise comparison randomness. In particular two Lac promoters contain a contiguous sequence of length 45 with 44 matches. This alignment, A10, is obtained for any gap weighting, w_k greater than or equal to unity. This alignment and its associated Sellers' distance of one is more than ten standard deviations away from the mean value of 22.0 among comparisons between shufflings of these two sequences! Thus while the proposed alignment, A11, keys on the "invariant Pribnow T" and/or the mRNA initiation site, the 19 matches are only what would be expected at random.

This last example points out the fact that even the use of keyed positions of functional equivalence (here the mRNA initiation site) cannot be used to define an optimal alignment without investigating the non-keyed similarity statistics first.

Discussion

The procedures described here are general in that they apply to any pair of linear sequences whose elements must remain in their original order but must otherwise be optimally aligned. For example, geological strata from different localities may be aligned to determine which strata are equivalent (Smith and Waterman 1980).

The procedures are also generalizable to higher dimensions, that is, it is possible to align three or more sequences simultaneously. This would seem desirable in view of the fact that the three optimal alignments of three sequences examined in all possible pairs need not give a set of mutually consistent gap placements. Our experience suggests, however, that even sequences of moderate length ($\leqslant 100$) require large amounts of computing space and time when done three dimensionally.

The procedures are further generalizable in that w_k need not be a linear but simply a monotonically increasing function of k. We generally use the form $w_k = w_g + kw_\ell$ where w_g is a gap penalty regardless of lengths and w_ℓ is a penalty on the length (k) of the gap, but one could use $w_k = w_g + w_\ell f(k)$ where f(k) is a polynomial in k. For alignments involving multiple sequences, this may involve gaps of different lengths (k and k') opposite each other (see Fitch and Yasunobu 1975), in which case the positive difference between w_k and w'_k should be used for the value of the gap.

Finally, one can generalize the weights by the addition of other weights. It has been observed that tandem repeats are frequently associated with the need for gaps as if tandem repeats facilitated unequal crossing over. One might therefore wish to reduce the value of w_k by some function of k (for example by kw_r where w_r was a constant less than w_ℓ) so that $w_k = w_g + kw_\ell - kw_r$ where $w_r = 0$ unless a tandem repeat is present.

Optimality is necessarily achieved by the algorithm but that should not mislead the user into thinking that the resulting alignment (or its significance) corresponds to some external reality independent of the weight assigned to the gap. Alignments A2, A3 and A4 are all optimal and make clear that a judicious choice of gap weights is required. One is constrained by the fact, for protein and nucleotide sequences, that the more distantly related the sequences are, the greater the distance between gaps must be to provide a sufficient number of paired elements to make that alignment significantly more similar than an equal number of pairs of randomly chosen elements. This suggests that one should then have a high gap weight to prevent an overabundance of gaps. On the other hand, the more distant two homologous sequences are, the more gaps its true history is likely to require and the smaller the likely interval between legitimate gaps on the average. This suggests a low gap weight. It is thus necessary to make a compromise between competing desiderata, a compromise that appears to be an uncertain function of the distance between the two sequences being compared. Apart from the two general rules given in the applications section, our experience suggests that for amino acid sequences, whose similarities are the maximum possible nucleotide matches for optimally chosen codons for the paired amino acid elements, a gap weight of $w_k \sim 3$ appears reasonable. For nucleotide sequences, a $w_k \sim 1.1$ appears reasonable.

It is not possible to avoid a decision on an appropriate gap weight. All alignments that are optimal by some objective criterion require a gap weight. To ignore gaps in determining the similarity (or distance) measure is in fact a decision to set $w_k = 0$.

There are values of w'_k that imply a negative value of w_k. If the gap weight for the distance measure were a constant, independent of gap length, the gap weight for the equivalent similarity measure would contain a term that decreased as a function of gap length, going negative for values of w_k for $k > 2w'_k$.

This curious, counter-intuitive relationship can be made to seem more reasonable if one considers the example of an alignment in which the m elements of the first sequence are aligned to a single gap of m null elements and, following the last element of the first sequence there is a second gap of n elements to which the elements of the second sequence are aligned. It would seem undesirable for such an alignment of two sequences to have any similarity, even if one is completely indifferent to the presence of gaps. If there is no deletion weight in a Needleman-Wunsch computation, $w_k = 0$, then the bracketed term of (4c) equals $(n + m)/2$ and the similarity takes the reasonable value of zero. That w_k should equal zero is not unreasonable since no amount of judicious gapping can increase the similarity beyond $(n + m)/2$ because the bracketed term of 4c is subtracted therefrom.

Now if $w_k = 0$ can be viewed as indifference to the presence of gaps, equation (5) implies that $w'_k = k/2$ is also indifferent. This value, used in equation (3) leads, for the alignment in the previous paragraph, to a distance $d = m + n$, again a reasonable value. Note that $w'_k = 0$ leads to a distance of zero for the same alignment, a most unreasonable result and indicates that values of $w'_k < k/2$ will tend to promote the introduction of gaps into an alignment.

These algorithms give two results, an alignment, with its measure, and, through scrambling of the sequences, an estimate of the significance of that alignment. A high significance only states that the resultant alignment is far from random, not that the proposed alignment is biologically meaningful. For example, setting $w_k = 0$ in a comparison of human alpha and beta hemoglobins gives a result of great significance. However, the alignment has many gaps in it with many amino acids in one sequence paired with amino acids in the other that no molecular biologist would agree

were pairs sharing a common ancestral codon. If $w_k = 3$ however, we get the presumably correct gaps in their correct places including, if terminal gaps are weighted, the gap of one residue between the first and second residue of the alpha hemoglobin. The significance of this alignment is also higher, but the point is that while a high significance implies that the sequences are similar, it does not, by itself, indicate that resultant alignment is itself optimal in some biological sense.

The problem of the interpretation of a significant similarity is more profound than a simple concern for a possible alternative alignment with still greater significance. It is a commonplace for investigators to conclude that significantly similar sequences are homologous which, since before Darwin's time, has meant that the two features (sequences in this case) shared a common ancestor. This ignores the possibility that the sequences are analogous and thus possess their similarity as a result of convergence. This latter alternative seemed unlikely in the past, but the present results on the lac promoter show it to be quite possible.

There is no way, in the face of alignment A10, that Rosenberg and Court's alignment A11, of two Lac promoters, can be said to be an homologous alignment. Nevertheless, there is reason to believe in the functional equivalence of the region from the proposed Pribnow box to the initiation site in the two sequences. Aligning them on that basis, in opposition to the more favorable alignment indicated in A10, gives 9 matches in the 13 nucleotides over that region, a rather significant result. There are only 10 additional matches in the remaining 44 nucleotides. To some degree, the result may be an accident in that the same sequence has two neighboring regions that are very promoter-like, A-T rich. But there is no evidence that either of these genes has two promoters operating and the T-A difference in the A10 homologous alignment is a change necessary to get the ninth nucleotide match in the A11 functional alignment. Thus, it appears that the significant sequence similarity shown in these promoter sites must be analogy, not homology. Clearly the observation of similarity and the inference of homology from that observed similarity should not be confounded by using homology to mean both the observation and the inference. It is equally clear from the lac promoter case that similarity is not per se sufficient to prove homology vis-a-vis analogy and that both homology and analogy occur in biological sequences, even in the same sequence. Worse yet, one cannot generally expect that the truly homologous sequence will be located near the analogous sequence and thereby rescue the investigator from a false inference.

It frequently arises that there is more than one alignment that yields the same optimal s or d value. These alignments usually have regions common to all alternatives as well as regions whose alignment is not unique. We are currently modifying our algorithm so as to identify these "locked" regions.

Acknowledgements. This work was partially supported by a grant to WMF from the National Science Foundation, DEB 78–14197.

References

Dayhoff MO (1978) Atlas of Protein Sequence and Structure, vol 5, suppl 3, Natl Biomed Res Found, Silver Springs, Maryland

Dickerson RE (1980a) Nature 280:210–211

Dickerson RE (1980b) Sci Am:242

Fitch WM (1977) Genetics 86:623–644

Fitch WM, Yasunobu KT (1975) J Mol Evol 5:1–24

Hamming RW (1950) Bell Syst Tech J 26:147

Needleman SB, Wunsch CD (1970) J Mol Biol 48:443–453

Pribnow D (1975) Proc Natl Acad Sci USA 72:784–789

Rosenberg M, Court D (1979) Annu Rev Genet 13:319–353

Rosenberg M, De Crombrugghe B, Busso R (1976) Proc Natl Acad Sci USA 73:717–721

Sellers PH (1974) SIAM J Appl Math 26:787–793

Sims J, Capor D, Dressler D (1979) J Biol Chem 254:12615–12628

Smith TF (1980) Mol Cell Biophys 1:3–14

Smith TF, Waterman MS (1980) J Geology 88:451–457

Waterman MS, Smith TF, Beyer WA (1976) Adv Math 20: 267–287

Received August 11, 1980/Revised May 8, 1981